Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

Sara Ahmadi

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no FP7-PEOPLE-2011-290000.

SIKS Dissertation Series No. 2017-39
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Cover design by Samaneh Karimi and Ehsan Ahmadi
Printed and bound by Ipskap Printing, Enschede

Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op donderdag 2 november 2017
om 16.30 uur precies

door
Sara Ahmadi
geboren op 5 november 1983
Tafresh, Iran

Promotoren:	prof. dr. Antal van den Bosch
	dr. Seyed Mohammad Ahadi
	(Amirkabir University of Technology, Iran)
Copromotoren:	dr. ir. Bert Cranen
	dr. Louis ten Bosch

Manuscriptcommissie:

prof. dr. Roeland van Hout (Voorzitter)	
prof. dr. Hamidreza Amindavar	Amirkabir University of Technology, Iran
prof. dr. Armin Kohlrausch	Technische Universiteit Eindhoven
dr. Ewen MacDonald	Danmarks Tekniske Universitet, Denemarken
prof. dr. ir. Hugo Van hamme	KU Leuven, België

Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

Doctoral Thesis

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defended in public on Thursday, November 2, 2017
at 16.30 hours

by
Sara Ahmadi
born on November 5, 1983
in Tafresh (Iran)

Supervisors:	prof. dr. Antal van den Bosch
	dr. Seyed Mohammad Ahadi
	(Amirkabir University of Technology, Iran)
Co-supervisors	dr. ir. Bert Cranen
	dr. Louis ten Bosch

Doctoral Thesis Committee:

prof. dr. Roeland van Hout (Chair)	
prof. dr. Hamidreza Amindavar	Amirkabir University of Technology, Iran
prof. dr. Armin Kohlrausch	Eindhoven University of Technology
dr. Ewen MacDonald	Technical University of Denmark, , Denmark
prof. dr. ir. Hugo Van hamme	KU Leuven, Belgium

Acknowledgements

This thesis is the result of my research in the Center for Language Studies (CLS) at Radboud university and the Speech Processing Research Lab (SPRL) at Amirkabir University of Technology. The work culminated in this thesis would not have been possible without the help and support of my supervisors. I learned a lot from and enjoyed working with my brilliant supervision team and I am thankful to them for many reasons. Bert, you are the reason I am here. My current life route would never have been the same if it was not because of the opportunity you gave me and the trust you put on me from the first day in CLS. I enjoyed working with you and discussing ideas with your sharp engineer mind. You were a very busy person who always had time for me. Antal, I appreciate your high level supervision and your excellent help in planning and organizing my work. Given my perfectionism, the thesis would never have been completed if it was not because of your encouragement and the right pressure you put on me in critical moments. Louis, dealing with your complicated comments and discussions was a challenge from which I learned a lot. Your extraordinary mathematical mind helped me view problems from different angles and it was a trigger to innovation. And my special gratitude goes to Lou. Dear Lou, you were always there for me, no matter if it was a weekend, late in the evening or holidays, I could always count on your help. I learned a lot from our discussions and also your constructive criticism. Your kind support and understanding in my stressful and frustrating moments was like a light in the darkness for me. I am also thankful to dr. Ahadi with whom I started my adventurous PhD. I appreciate his help and guidance into speech processing research which was an unknown area to me before. I also appreciate his understanding when I decided to move from his lab in order to go for my ambitions and experience more.

During my PhD, I was involved in the initial training network INSPIRE where I met senior researchers and benefited from their ideas and comments among which I am especially grateful to Hugo Van Hamme, Tuomas Virtanen and Martin Cooke. I also would like to express my gratitude towards Torsten Dau for being my host during my secondment at DTU. The discussions with him were very helpful in designing part of my experiments and interpreting the results. I am also grateful to Tobias May for his collaboration and advice during my secondment at DTU. Being part of the INSPIRE network, I found the chance to meet the amazing

INSPIRE fellows, young researchers with whom I collaborated and shared lots of nice memories. Alex, Cezara, Deepak, Dorina, Gustav, Huarda, Johannes, Juliane, Kurt, Nemanja, Ricard and Tom, I am very glad to get to know you all. Moreover, I thank Deepak for his help in developing my large vocabulary recognizer using Kaldi.

CLST was a very pleasant work environment for me. I would like to thank the director of CLST, Henk who was also a very supportive colleague. The lovely secretary of CLST, Hella was one of the first people I met at Radboud University. She was so nice and friendly and she made me feel at home. I would like to thank my colleagues and friends in the Erasmus building who made my time joyful. Polina, Jule, Mario, Eric, Vanja, Steve, Michele, Job, Remy, Marjoke, Wessel, Claire, Ferdy and Hamed, you are a great group of people I was lucky enough to meet and spend time with you during coffee breaks, lunch and our occasional activities outside our work place.

I started my life in the Netherlands with a very warm welcome from Els, my lovely landlady with a big heart, who provided a calm and cozy home for me. I would like to thank her for her emotional support.

Last but not least, I would like to express my deepest gratitude to the two angels of my life, my parents, Ashraf and Mahmoud; I love you two so much and I am grateful for all your support and encouragement especially during my adventurous study life. I also thank my brothers, Mehdi and Ehsan for being an endless source of emotional energy for me.

Contents

Chapter 1

Introduction

Automatic speech recognition (ASR) refers to processes that enable machines to map speech signals into sequences of words. A typical ASR system consists of three main building blocks. First, a frontend module that converts the analogue speech signal into a sequence of acoustic feature vectors. Second, a module that converts the sequences of feature vectors into posterior probabilities (or likelihoods) of a potentially large number of sub-word units. This provides a lattice which is a description of how the probability of sub-word units change over time. And third, a backend decoding module that searches for the most probable path through the lattice, resulting in the most likely sequence of words. With such an architecture state-of-the-art ASR systems can achieve recognition accuracies that are close to human performance, at least in conditions where the speech signal is recorded in a quiet environment. In comparison to human listeners, however, under more adverse conditions encompassing the presence of environmental noise, competing talker(s), room reverberation, channel distortion, or a combination thereof, performance of ASR systems drops more severely and more rapidly.

In natural-world situations, sounds are usually interleaved and overlapped in time and their components are interleaved and overlapped in frequency. Studies on how humans separate individual sound sources in natural-world situations, i.e., how they perform auditory scene analysis (ASA), indicate that regularities in both time and frequency play an important role that cause certain sound sound components to be perceived as coming from a single source (grouping) or from different sources (segregation) (Bregman, 1994). Also, listening experiments in which

listeners are exposed to artificially corrupted speech signals of which substantial parts are missing or are masked by noise, revealed that the human auditory system is able to exploit many different types of cues to recognize the target speech to maintain quite high recognition rates (Drullman, 1995; van Dijkhuizen et al., 1991; Festen and Plomp, 1990). Humans even outperform ASR systems in tasks where semantic predictability is absent, such as in recognizing digit sequences (Meyer, 2013) or phonemes (Meyer et al., 2011).

The main question we address in this thesis is to what extent the performance of ASR in noisy conditions can be improved by more explicitly incorporating knowledge about human speech perception in noise.

Over the past 30 years, a large number of methods were proposed and published in the context of noise robust ASR (Li et al., 2014). Noise robustness can be achieved either by adding modules to the system, such as modules for noise reduction (Boll, 1979; Lim and Oppenheim, 1979) and feature enhancement (Stouten et al., 2006; Raj et al., 1998, 2004), or by modifying each of the modules to be more noise robust. Examples of the second group of methods include modified or alternative feature extraction (Hermansky, 1990; Kim and Stern, 2010; Moritz et al., 2011; Fazel and Chakrabartty, 2012) and noise robust probability estimation (Cui and Gong, 2007; Gemmeke et al., 2011b; Hurmalainen et al., 2011a; Hinton et al., 2012).

In this thesis three novel approaches are explored. The first, which we will introduce in Section 1.1, can be considered a feature enhancement technique and in the other two we propose alternative modules for feature extraction (see Section 1.2) and probability estimation (see Section 1.3) respectively. In all approaches proposed in this thesis, we consider sparse coding as an effective tool to deal with signals of which only a small portion is available in an undisturbed form. Such circumstances are typical for processing speech that is heavily corrupted with noise: when a significant part of the energy in the time-frequency plane of the signal is dominated by noise, only a limited part of the original speech signal will stand a chance to remain (relatively) unaffected. In neurobiology, sparse coding refers to the representation of objects in the form of a strong activation of a relatively small set of neurons in the visual (Willmore and Tolhurst, 2001; Olshausen and Field, 1997) or auditory (Hromádka et al., 2008) cortex of a human (or more

generally, mammalian) brain. In machine learning, sparse coding denotes representing data vectors as a sparse linear combination of basis elements (Mallat, 1999; Mairal et al., 2008b). The basis elements can be analytic functions such as wavelets (Mallat, 1999), basis vectors learned from data (Mairal et al., 2009a; Aharon et al., 2006) or exemplars (Gemmeke et al., 2011b). Our motivation to use sparse coding based techniques, is that according to the compressive sensing (CS) theory (from the field of signal processing), a sufficiently accurate reconstruction of a signal is possible using the sparse codes computed from a relatively small portion of the original signal (Candes and Wakin, 2008). Therefore these techniques seem suited par excellence for recognizing words in auditory scenes where the target speech is obscured by other sound sources.

1.1 Feature enhancement using missing feature imputation

In the first study, we focus on compensating the effect of noise on feature vectors of the noisy speech signal. When a noisy speech signal is represented as a power spectrogram, i.e., a spectro-temporal matrix representation depicting the acoustic power of individual time-frequency elements, each element can be considered as the sum of the power of underlying speech and background noise. Since the acoustic energy of speech and noise will typically be distributed differently over the time-frequency plane, there are regions in the spectrogram that are dominated by noise while other regions can still be considered as very similar to (i.e. a reliable estimate of) the speech power. The noise-dominated regions are considered as missing data and need to be treated differently from the undisturbed regions. This is called the missing feature (data) problem for which the solution is addressed in a group of techniques called missing data techniques (MDT) (Cooke et al., 1997; Raj, 2000). MDT approaches for speech recognition can be divided into two categories: marginalization (Cooke et al., 2001; Barker et al., 2001) and imputation (Raj et al., 1998, 2004). Missing feature imputation techniques aim at reconstructing the noise-dominated elements in the spectrogram using the reliable speech components and hinges on exploiting the inherent redundancy in the speech signals. In this thesis we propose a missing feature imputation approach using manifold based compressive sensing.

In speech, like in many natural signals, the number of factors causing perceptually relevant differences are considered as degrees of freedom and are usually far fewer than the number of embedding dimensions (i.e. the number of acoustic features that are observed at each time instant) (Tenenbaum et al., 2000; Tosic and Frossard, 2011). The degrees of freedom for speech are determined by the articulatory constraints of the human speech production system (Jansen and Niyogi, 2013) and the constraints that follow from the limited number of distinctive speech sounds of a language (Stevens, 2000). Therefore, it may be expected that a set of particular speech signals resides on a nonlinear manifold of which the effective dimensionality is much lower than the signal space. In cognitive science, certain findings are sometimes interpreted as indications that the human brain also relies on these intrinsic nonlinear manifolds to perceive constancy in perceptual stimuli, such as images, even though the raw sensory input may be quite variable. It has been hypothesized that manifolds are stored in the brain as manifolds of stable neural activity patterns (Seung and Lee, 2000). This leads us to the first research question of this thesis.

RQ1: To what extent can we benefit from the underlying manifold of speech to reconstruct the missing parts of an incomplete speech spectrogram?

1.2 Human inspired feature extraction

Our motivation for the second study presented in this thesis was to benefit from the extensive knowledge about the human auditory system and the way in which humans cope with noise. By doing so, we wanted to narrow (or at least better understand) the gap between human and ASR performance in noise. Although, admittedly, there is no empirical evidence about the ways in which sequences of tonotopic representations are used in the brain as an interface to a putative mental lexicon, it is conjectured that the mapping of concrete tonotopic representations or more generalized (or abstract) representations is mediated by an exemplar-based procedure (Goldinger, 1998). Following this idea, we suggest to use an approach reminiscent of exemplar-based sparse coding presented in Gemmeke et al. (2011b) which has already been used successfully for noise-robust ASR on the conventional Mel spectral features. Sparse coding is able to handle very high dimensional and

redundant feature representations without the need for dimensionality reduction procedures that would most likely hamper the distinction between features that are dominated either by noise or speech. This property enables us to explore the possibility of estimating sub-word posterior probabilities from feature representations that have close connections to models of the human auditory system. One of such models, proposed in Jørgensen and Dau (2013); Jørgensen et al. (2013), has shown to be quite effective in predicting the intelligibility of speech signals in many different noisy situations. The model tries to capture the most salient properties of the peripheral human auditory system. First, it models the frequency analysis of the speech signal carried out by the basilar membrane of the cochlea. As a second step, it models the changes over time of the neural responses of the hair cells of a certain frequency region. The output of the model is the so-called modulation spectrum and provides a rich and highly redundant feature space to explore. While the intelligibility predictors in Jørgensen and Dau (2011) rely only on the envelope signal to noise ratio (i.e. relatively long-term averages of speech and noise powers), we are interested in the requirements that are posed when the goal is to extract all information along the time dimension encoded in the modulation spectrum and that humans might use to understand the message in detail. The second research question we address in this thesis is:

RQ2: Since modulation spectrum features appear to be a suitable basis for predicting intelligibility of speech in noise, to what extent are those features also an adequate starting point for a sparse coding based ASR system that is robust against noise?

1.3 Supervised dictionary learning for sparse classification of speech units

In the study introduced in Section 1.2 and described in detail in Chapters 3 and 4, we employed an exemplar-based sparse coding procedure to estimate the posterior probability of sub-word units. Exemplar-based sparse coding uses a collection of speech segments, called exemplars, as the sparsifying basis. As the variations in the data set increases, the number of exemplars required to cover the entire signal space will be higher. This motivated us to try and replace the exemplar dictionary by a learned dictionary (Tosic and Frossard, 2011) in which a limited

number of atoms are trained using a large number of speech exemplars covering the whole signal space. The dictionary learning methods are designed with the primitive aim of minimizing the so called reconstruction error (i.e. the mean squared distance between original and reconstructed signal using sparse codes) (Aharon et al., 2005; Mairal et al., 2009a, 2008b). Although it is possible to use such learned dictionaries to estimate the posterior probabilities, they might be sub-optimal for that purpose. The dictionary learning formulation enables us to include a mathematical extension to modify and improve the discriminative power of the learned dictionary (Jiang et al., 2013; Yang et al., 2011; Mairal et al., 2008a). To effectively use learned dictionaries in a sparse coding based ASR, we need to adapt the existing dictionary learning methods to the posterior probability estimation problem. Therefore, the third research question we address in this thesis is

RQ3-a: To what extent can we improve the accuracy of the estimated posterior probabilities of speech units by customizing the learning procedure in such a way that the probabilistic classification accuracy is also considered as an objective during learning?

RQ3-b: To what extent does posterior probability estimation based on learned dictionaries increase noise robustness in the sparse coding based AR system?

1.4 Overview

In this section we provide a short overview of the four chapters included in the body of this thesis. We then finish the thesis by summarizing our findings, reflecting on them in view of our research questions, and offering our recommendations for future work.

1.4.1 Chapter 2: Missing Feature Imputation Using Manifold-based Compressive Sensing for Noise-Robust Speech Recognition

In Chapter 2, we address the idea raised in Section 1.1. We propose a novel missing feature imputation approach based on a combination of manifold learning and

compressive sensing. We consider segments of speech spectrograms as the embedding space and use a mixture of linear probabilistic principal component analysers to model the low-dimensional nonlinear manifold that supports the speech signals. We then use the reliable spectro-temporal elements in a noise-corrupted spectrogram to locate the signal on the manifold. This allows reconstructing the complete spectrogram, that is subsequently transformed into conventional MFCC features for use with a conventional HMM decoder as well as a modern DNN decoder. The method is examined on both small and medium size vocabulary data sets (Aurora2 and Aurora4).

The chapter is adapted from: Ahmadi, S., Ahadi, S.M., ten Bosch, L. & Boves, L.W.J. Missing Feature Imputation Using Manifold-based Compressive Sensing for Noise-Robust Speech Recognition. Manuscript submitted to *Circuits, Systems, and Signal Processing* in March 2016. Revised and resubmitted in May 2017.

1.4.2 Chapter 3: Sparse coding of the modulation spectrum for noise-robust automatic speech recognition

Chapter 3 is the first step in implementing the idea raised in Section 1.2. In this article, we propose a sparse coding based speech recognition system that uses exemplars consisting of modulation spectrum features. We assume that the multiple time resolutions with which the energy in a certain frequency region is represented in modulation spectra (modelling the first neural processing stages of the hair cell outputs) is key to providing sufficient redundancy to let a sparse coding based recognition system benefit from it and become noise robust. As long as the high-dimensional modulation spectrum contains enough features that are not affected by the noise, they will dominate the distance measure in a sparse coding engine. The feature extraction module, adopted from a speech intelligibility prediction study, is adapted to the recognition task. The required connecting modules between the feature extraction and posterior probability estimation are designed to match mathematical and numerical constraints. Experiments using a small vocabulary task show that the novel approach is promising and it also reveals important considerations to take into account in order to improve the performance of such a system.

The chapter is published as: Ahmadi, S., Ahadi, S.M., Cranen, B. & Boves, L.W.J. , Sparse coding of the modulation spectrum for noise-robust automatic speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014, 2014(1): 1-20.

1.4.3 Chapter 4: Human-inspired Modulation Frequency Features for Noise-robust ASR

Continuing the study on human inspired feature extraction, we take the recognition system designed in Chapter 3 and focus on the details of the feature extraction procedure. By varying various design parameters, we aim at increasing the performance of the whole system as well as achieving a better understanding of the way in which details in the representation of the modulation spectra impact the resulting recognition accuracy. We also investigate which factors play an essential role in increasing noise robustness and compare those findings with the existing evidence about how human listeners perform in very noisy conditions. Based on the insight gained in this study, we arrive at ideas that can also be beneficial in improving existing classical ASR systems.

The chapter is published as: Ahmadi, S., Cranen, B., Boves, L., ten Bosch, L., & van den Bosch, A. Human-inspired Modulation Frequency Features for Noise-robust ASR. *Speech Communication* (2016).

1.4.4 Chapter 5: Class-Likelihood-Consistent Dictionary Learning for Probabilistic Classification

To address the questions raised in Section 1.3, we designed a speech unit posterior probability estimation procedure using dictionary learning methods. The estimation of speech unit probabilities can be considered as a probabilistic classification problem for which there are approaches proposed in the machine learning and pattern recognition literature. In the 5th chapter of this thesis, we propose a dictionary learning based probabilistic classifier and develop a mathematical extension to the existing dictionary learning procedures to improve its probability estimation accuracy. In doing so, we assume that reference class likelihood vectors for the training observations are available, in which the vector elements represent

the class membership probabilities for that particular observation. The aim of the proposed learning method is to decrease the difference between these reference likelihoods on the one hand, and the ones estimated by the proposed probabilistic classifier on the other. We evaluate the proposed method by (1) examining the classification performance on a set of synthetic data specifically designed to show the strengths and weaknesses of the method, and (2) investigation of the estimated posterior probabilities of sub-word units in a real ASR task. We also investigate how effective the proposed method is in estimating sub-word unit posterior probabilities when applied to noisy speech. The recognition accuracy obtained when the estimated probabilities for noisy speech are fed into a Viterbi decoder is computed. We compare the recognition accuracy at various SNR levels with the results obtained using the exemplar based ASR reported in Chapter 4.

Part of this chapter is submitted as: Ahmadi, S., Cranen, B., ten Bosch, L., Boves, L., Class-Likelihood-Consistent Dictionary Learning for Probabilistic Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Chapter 2

Missing Feature Imputation Using Manifold-based Compressive Sensing

Missing feature imputation is an approach to noise-robust automatic speech recognition that tries to reconstruct the noise-corrupted components of speech signals prior to recognition. The estimation of the missing part of the spectrogram exploits the inherent redundancy in the spectrographic representations of speech signals. In this chapter, a novel missing feature imputation approach based on a combination of manifold learning, using a method that has not been applied to speech before, and compressive sensing is presented. Segments of spectrograms are considered as spanning the embedding space. The low-dimensional non-linear manifold that supports the speech signals is approximated by means of a mixture of probabilistic Principal Component Analyzers. The location of a noise-corrupted spectrogram segment on the manifold can be found using only the reliable spectro-temporal features. This allows reconstructing the missing parts of the spectrogram. We present a novel, computationally efficient closed form for the reconstruction procedure. The reconstructed log-Mel spectra are transformed into conventional Mel-frequency cepstral coefficients for use with a conventional Gaussian mixture model – hidden Markov model back-end decoder. Alternatively, the reconstructed spectra can be used as input to a back-end that uses a Deep Neural Network in combination with a finite state transducer. Experiments using

the AURORA-2 and the AURORA-4 corpora show that the novel feature imputation method compares favorably against competing imputation techniques.

2.1 Introduction

Real-world applications of automatic speech recognition (ASR) must deal with signals that are affected by various kinds of distortions. Addition of background noise is among the most frequent distortions. As the signal-to-noise ratio (SNR) decreases, ASR performance degrades much faster than human speech recognition (Lippmann, 1996; Meyer, 2013). The apparent difficulty of coping with noise in ASR has led to a number of different approaches in a large literature that is summarized and reviewed in Acero (1993); Kolossa and Haeb-Umbach (2011); Virtanen et al. (2012). This chapter introduces a novel approach to noise robust ASR based on the missing data technique (MDT). MDT approaches assume that it is possible to identify the noise-dominated elements of the speech signal representation and consider them as 'missing' features. The aim of MDT is to handle the missing features using the features that are present based on the insight that in many applications the features used to represent physical processes are not statistically independent. A recent overview paper (Li et al., 2014), that cites a very large number of papers and books, proposes a taxonomy of approaches to noise-robust ASR based on the ways in which the intractable mathematics of modeling the statistical distributions of the acoustic features of noisy speech is approached. The taxonomy comprises five –not necessarily independent– dimensions: (1) does an approach operate in the feature domain; (2) does it use prior knowledge about corrupting noises; (3) does it explicitly model the distortion caused by the noise; (4) is it based on uncertainty processing; and (5) does it involve training a joint model of speech and noise? In the taxonomy of Li et al. (2014) MDT is classified as $\{yes, no, no, no, yes\}$.

This chapter presents a novel implementation of MDT. The literature on MDT distinguishes two major approaches (Raj and Stern, 2005), namely classifier modification (e.g. Cooke et al., 2001) and data imputation (e.g. Raj et al., 2004). This chapter introduces a novel approach to data imputation. The novelty is that the imputation approach is based on combination of two signal processing concepts that were recently introduced in the speech processing community: compressive

sensing (CS) and manifold learning. Compressive sensing (e.g. Candes and Wakin, 2008; Gemmeke et al., 2010, 2011b; Ahmadi et al., 2014) holds that signals can be recovered (almost) perfectly from fewer measurements than the Nyquist rate, if the projection of the signal in some –possibly over-complete– basis is sparse. Manifold learning holds that the 'configuration space' or 'latent space' of many real-world events is characterized by a much smaller number of degrees of freedom than the dimensionality of the features used to represent signals in the observation space (Tenenbaum et al., 2000; Jansen and Niyogi, 2013).

Most of the information in speech signals is encoded in the dynamic changes of the spectrum over time, rather than in static short-time spectra (Mlouka and Liénard, 1974; Furui, 1981). Conventionally, the dynamics is represented by appending the first and second time derivative to the static spectral features, or by using stacks of consecutive feature frames. This leads to feature vectors with typically a dimension of at least 39. However, it has long been known that no more than about 15 parameters are sufficient for controlling an articulatory speech synthesizer (Rubin et al., 1981). It has been hypothesized that articulatory features can also be used for automatic speech recognition. However, attempts to develop ASR systems that use articulatory features have met with substantial limitations (e.g. Ghosh and Narayanan, 2011), most probably due to the non-linear and non-unique mapping from acoustics to articulation (Najnin and Banerjee, 2015). Due to the overwhelming success of acoustic modeling based on Gaussian mixture models (GMM) and artificial neural nets (ANN) in speech processing, approaches based on manifold learning have obtained much less attention in the speech community than in the image processing community. Recently, approaches to manifold learning that do not try to interpret representations in the latent space in terms of phonetic features, have shown to be useful in speech recognition (Jansen and Niyogi, 2013; Huang et al., 2016). In this chapter, we propose to reconstruct the missing (noise-dominated) features of speech using the underlying manifold. Therefore we establish a novel way of looking at the missing data problem and it is likely to result in an approach that can reconstruct the missing data using very few reliable features.

From the image processing field it is known that non-linear manifolds can be approximated by means of a mixture of locally linear probabilistic Principal Component Analyzers (MPPCA) (Bishop and Winn, 2000; Baraniuk and Wakin, 2009; Wakin, 2010; Chen et al., 2010). Bishop in chapter 12 of Bishop (2006) has shown

that it can be guaranteed that the representation of the original data on the manifold is sparse, if suitable prior distributions are imposed on the parameters of the sub-models. Our work is the first attempt to apply MPPCA modeling to noise-robust ASR.

The representation of speech signals in the form of spectrograms (three-dimensional pictures with time and frequency on the axes and power displayed as grey level or colors) suggests a parallel with image processing. Therefore, we reasoned that it is possible to model non-linear speech manifolds as a mixture of probabilistic PCAs. However, the challenges facing image reconstruction and classification are very different from the challenges that must be overcome in noise-robust ASR. An important difference between image processing applications, such as inpainting, and MDT for noise robust ASR is that, more often than not, missing speech features do not form connected patches. On the contrary, missing features appear to be scattered. Because of this scattering it is more difficult to determine which features are missing. Another difference is at the application level. In image restoration, and also in noise suppression, either the reconstruction is the final output and evaluation is left to intelligent human processing or the reconstructed image is used for image classification. In noise-robust ASR, however, neither the reconstruction nor the classification result is the final output, but the input to a back-end processor that implements a constrained search for the most likely solution. It is theoretically and practically very difficult to use properties of the search algorithm in the back-end to infer ways in which the observation features can be improved.

The guarantee that the representation on a manifold learned by means of MPPCA is sparse provides a natural link with compressive sensing. In previous attempts to use the CS-approach to robust ASR proposed by Gemmeke et al. (2011b) an over-complete embedding space was created in the form of a dictionary of $8,000$ exemplars that are semi-randomly selected from a training corpus. Their exemplars consisted of stacks of 26-dimensional Mel-spectrum frames; the stacks could become as large as 30 frames. The dictionary used in Ahmadi et al. (2014) consisted of $17,000$ 105-dimensional frames of modulation spectra. However, if we use the latent space formed by means of MPPCA as the space in which we operate, we can consider the representation in the form of raw features as representations in the embedding space (Baraniuk and Wakin, 2009). This eliminates the need for

artificially constructing an over-complete embedding space. Thanks to the sparseness of the representations on the manifold, it is possible to obtain an accurate mapping on the manifolds, despite the fact that not all features in the embedding space are available. Although the mapping from the embedding to the latent space is many-to-one, an inverse mapping from the latent to the embedding space can be defined that yields vectors in the observation space that are sufficiently similar to the original observations (c.f. Bishop (2006) Chapter 12 on *Non-linear Latent Variable Models*).

In previous applications of CS to speech processing reconstructions derived from the optimal representation in the embedding space were obtained by means of techniques based on Lasso (Friedman et al., 2001) or non-negative matrix factorization (Lee and Seung, 1999). The reconstruction of full feature vectors on the basis of only the subset of available features is the most computation-intensive part in an actual application. For the reconstruction using MPPCA we develop a novel, computationally efficient, closed-form procedure, which we consider as another contribution of our work. The results of our experiments show that the novel reconstruction procedure is viable.

We will apply manifold-based compressive sensing for missing feature imputation to two well-known robust ASR tasks, AURORA-2 (Hirsch and Pearce, 2000) and AURORA-4 (Parihar and Picone, 2002; Parihar et al., 2004); both corpora consist of speech that was artificially corrupted by adding noise. Since both the clean and the noisified speech are available, it is possible to create a so-called oracle mask, i.e., a mask that is free of mask estimation errors. However, it should be noted that the performance loss due to mask estimation errors is also an important criterion in evaluating missing data imputation techniques.

The remainder of this chapter is structured as follows. Section 2.2 explains how the mixture of probabilistic PCA analyzers can be used for estimating the non-linear manifold that supports speech signals. Also, our efficient proposed procedure for imputing missing features from a sub-set of reliable features on the manifold is explained. In Section 2.3 an overview of the experimental setup is described followed by the details and results of the two experiments. Sub-sections 2.3.2 and 2.3.3 are devoted to a small and a medium-size vocabulary noise-robust ASR task. Section 2.4 discusses the results of the two experiments in the framework of noise-robust ASR, and Section 2.5 summarizes the main conclusions of the research.

2.2 Method

2.2.1 Modeling the speech manifold

Most previous speech missing data imputation methods which are based on compressive sensing create an overcomplete basis (spanning the embedding space) by selecting a very large number of (labeled) vectors from the high-dimensional space in which raw and uncorrupted measurements were made (Gemmeke et al., 2010, 2011a). This over-complete basis is then used for reconstructing new, possibly corrupted measurements. In the manifold-based CS technique, which we use in our proposed MDT approach, there is no need for constructing an overcomplete basis. Rather, we consider the raw measurements as representations in an embedding space, and apply manifold learning to find the low-dimensional latent space. We will approximate the non-linear manifold that supports the clean speech representations by means of a mixture of locally linear PCA sub-models (Tipping and Bishop, 1999a).

2.2.1.1 The MPPCA Model

Principal Component Analysis (PCA) is a well-known technique for finding a low-dimensional basis that can be considered to span the latent space for high-dimensional measurements. In Tipping and Bishop (1999b); Roweis (1998); Bishop and Winn (2000) it was shown that classical PCA can be considered as a limiting case of a Gaussian latent variable model. This turns classical PCA into probabilistic PCA (PPCA) (Tipping and Bishop, 1999b). Modeling a data set with (P)PCA invokes the assumptions that the distribution of the data is homoscedastic. It has been observed many times that real data, including speech data, do not adhere to homoscedastic distributions (e.g. Chang and Glass, 2007; Huang et al., 2011). This observation raised the question whether it is possible to apply (P)PCA to subsets of a complete data set that are reasonably homoscedastic, and then combine the resulting sub-models to obtain a complete representation. PPCA yields to probabilistic outputs, and these can be merged in a mixture of probabilistic principal component analyzers (Tipping and Bishop, 1999a). Using the resulting mixture of probabilistic PCA (MPPCA) model, a piecewise linear approximation

of a nonlinear manifold can be achieved. The generative equation of the MPPCA model is

$$\vec{t_n} = \vec{W_m}\vec{x_n} + \vec{\mu_m} + \vec{\varepsilon} \tag{2.1}$$

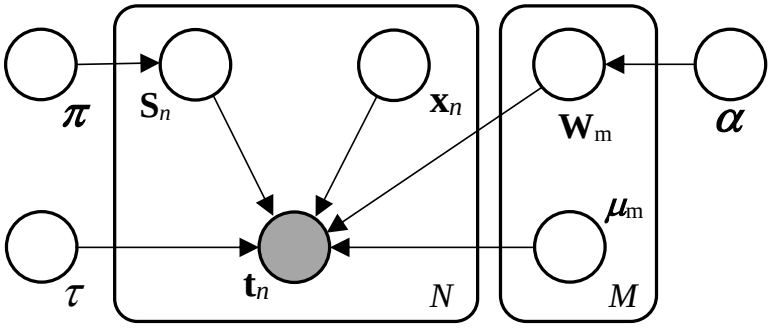


FIGURE 2.1: Representation of the MPPCA model as a probabilistic graphical model. The graph shows the hierarchical prior over **W** governed by the shared parameter vector α. The left box denotes the N independent observations $\vec{t_n}$ together with the corresponding latent variables $\vec{x_n}$ and $\vec{S_n}$. The right box denotes the M sets of parameters associated with each mixture component.

The corresponding graphical model is shown in Figure 2.1. The d-dimensional observed variables $\vec{t_n}$ $(1 < n < N)$ in the left-hand box can be represented by a q-dimensional latent variable $\vec{x_n}$ with zero-mean and unit variance Gaussian distribution $P(\vec{x}) = \mathcal{N}(0, \vec{I_q})$, where $\vec{I_q}$ is the q-dimensional unit matrix. The index m refers to the m^{th} PPCA sub-model (mixture component). The variables in the right-hand box ($\vec{W}$ and $\vec{\mu}$) are different for each sub-model. Each $\vec{W_m}$ is a $d \times q$ $(q < d)$ matrix, the columns $(\vec{w_{m,i}})$ of which are the transformation kernels. The expectation of $\vec{t_n}$ will be $\vec{\mu_m}$, which is a d-dimensional vector. $\vec{\varepsilon}$ is a noise vector (i.e, modeling noise) with zero-mean Gaussian distribution and covariance $\tau^{-1}\vec{I_d}$ (τ denotes the precision). To control the effective dimensionality of the latent space, Bishop and Winn (2000) proposed to introduce a hierarchical prior over the matrix $\vec{W_m}$ in (2.1). The columns of $\vec{W_m}$ are assumed to have conditional Gaussian distribution governed by the precision vector of $\vec{\alpha}$, where $\vec{\alpha} = \{\alpha_1, \cdots, \alpha_q\}$ is a random variable with Gamma distribution. Each α_i controls the inverse variance of corresponding $\vec{w_{m,i}}$. Therefore, if the mean of the posterior distribution of a particular α_i has a large value, this means that the corresponding $\vec{w_{m,i}}$ has small variance and that this direction has less importance for the data representation in the latent space. Thus it can be ignored without significant loss

in information. Hence the proper dimensions of the latent space (q) can be chosen using the values computed for $\vec{\alpha}$ in the training phase. It is also necessary to define prior distributions for the random variables $\vec{\mu_m}$ and τ. To guarantee that the computation of the posterior distributions is tractable the priors of $\vec{\mu_m}$ and τ are chosen to have Gaussian and Gamma distributions, respectively.

To enable a probabilistic selection of the optimal mixture component in the (re)construction of an observation $\vec{t_n}$, an M-dimensional binary latent variable $\vec{S_n}$ is defined in which all components are *zero* except the one which indicates the mixture component involved in construction of $\vec{t_n}$. $\vec{S_n}$ is a random vector with multinomial distribution with parameters $\vec{\pi} = \{\pi_1, \cdots, \pi_M\}$, where π_m is the prior probability that the m^{th} analyzer accounts best for the data. Hence, S_{mn} is the m^{th} entry of $\vec{S_n}$ and

$$\begin{cases} P(S_{mn} = 1) = \pi_m \\ P(S_{mn} = 0) = 1 - \pi_m \end{cases}$$

Finally, $\vec{\pi}$ is considered as a random variable having a Dirichlet distribution with constant parameters (Bishop and Winn, 2000).

2.2.1.2 Model Training

During model training the parameters $(\vec{W}, \vec{\mu}, \tau, \vec{\alpha}, \vec{\pi})$ and latent variables $(\vec{X_n}, \vec{S_n})$ must be estimated on the basis of a set of observations in a training corpus. $\vec{W}$ and $\vec{\mu}$ are the sets of all $\vec{W_m}$ and $\vec{\mu}_m$ corresponding to sub-models. Each $\vec{t_n}$ corresponds to a unique $\vec{W_m}$ and $\vec{\mu}_m$. In our problem, observations are spectrographic representations of speech utterances. Given $\vec{D}$, a $d \times N$ matrix of N independent observations $\vec{t_n}$, the MAP estimate of the unknown set $\psi = \{\vec{W}, \vec{\mu}, \tau, \vec{\alpha}, \vec{\pi}; \vec{X_n}, \vec{S_n}\}$ is obtained by simultaneously maximizing the a posteriori probability of the unknown parameters and latent variables $(P(\psi|\vec{D}))$ given the training data.

Computing the a posteriori distribution $P(\psi|\vec{D})$ involves a complex integral, which is computationally intractable. Fortunately, there are at least two different approaches to obtain an approximate solution (Bishop, 2006):

1. Approximate inference methods, such as variational inference (Bishop and Winn, 2000), which make use of some simplifying assumptions to make the problem tractable.

2. Sampling methods, such as Gibbs sampling (Geman and Geman, 1984), in which unknown parameters are estimated by generating a number of samples in the complex a-posteriori pdf.

In our proposed method, we used the Gibbs sampling method to train the model that approximates the manifold. For that purpose we derived the a-posteriori probability distribution function of each unknown given the observations and other unknowns, i.e. $P(\vec{W}|D, \vec{\mu}, \tau, \vec{\alpha}, \vec{\pi}; \vec{X}_n, \vec{S}_n)$, $P(\vec{\mu}|D, \vec{W}, \tau, \vec{\alpha}, \vec{\pi}; \vec{X}_n, \vec{S}_n)$, $P(\tau|D, \vec{W}, \vec{\mu}, \tau, \vec{\alpha}, \vec{\pi}; \vec{X}_n, \vec{S}_n)$, and so on. These probability distribution functions are then used to generate samples. The training procedure can be summarized as follows:

- Initialization: all the unknowns in ψ are initialized by a random value derived from their prior distributions (cf. section 2.2.1.1) .
- Gibbs iterations: in each iteration, one sample of each of the unknowns is generated from their a-posteriori distribution, with all other unknowns set to the value in the previous iteration.
- Averaging: after a large number of Gibbs iterations, the average value of the samples of each unknown (leaving out the first 30) is computed and taken as the final estimate for that unknown.

To obtain a visual impression of the convergence of the Gibbs procedure, the intermediate estimates of the unknown parameters were used to reconstruct the training data and compute the average reconstruction error. We observed that 600 to 1000 Gibbs iterations were sufficient to obtain a stable set of parameters.

To be able to perform a training, the definition of the input features, the number of dimensions q to retain in the PCA analyzers and the number of mixture components M must be fixed. As explained in Section 2.2.1.1, the proper value of q can be derived from the values of the elements in the vector $\boldsymbol{\alpha}$. The definition of the features and the selection of the optimal value of M are task-dependent; these issues will be treated in Section 2.3. In brief: a large number of models with different values of these two parameters were trained, and the eventual recognition accuracy on a development set was used to find the optima.

After the training phase, the estimated values for $\vec{\mu}_m$ determine the mean of each MPPCA sub-model in the mixture, and the columns of $\vec{W}_m$ define the proper local coordinate system. Actually, $\vec{W}_m$ is a linear transformation that leads to the most sparse representation. The estimated values of $\vec{\pi}$ indicate the amount of contribution of each mixture component to the data space. In other words, π_m provides a way to know what fraction of data is modeled by the m^{th} PPCA sub-model in the m^{th} plane. $\vec{\alpha}$ and τ are also estimated in the training phase, as well as the latent variables for each observation. $\vec{S}_n$ denotes the active mixture component involved in the construction of $\vec{t}_n$, while $\vec{x}_n$ locates the representation of $\vec{t}_n$ in the indicated hyperplane.

2.2.2 Recovering missing data in the reconstruction phase

In the reconstruction phase, the most probable representation of the complete observations must be computed from the subset of the reliable features. For this purpose, the MPPCA model obtained during the training will be used. In the reconstruction phase, the model parameters $\{\vec{\mu}, \tau, \vec{\alpha}, \vec{\pi}, \vec{W}\}$, which were random variables in the training phase, are treated as fixed parameters. The dimension of the sub-models (q) and the number of mixture components (M) are also known. However, even with these parameters fixed, reconstruction is not completely straightforward.

Suppose that d_r components of a spectrogram are available as reliable features and the remaining $d_u = d - d_r$ features are missing (because these are dominated by the noise). In order to reconstruct the missing components, the representation of the signal in the q-dimensional space must be estimated. To formulate the reconstruction process, equation (2.1) is broken into

$$\begin{cases} \vec{t}_u = \vec{W}_{m,u}\vec{x} + \vec{\mu}_{m,u} + \varepsilon \\ \vec{t}_r = \vec{W}_{m,r}\vec{x} + \vec{\mu}_{m,r} + \varepsilon \end{cases} \tag{2.2}$$

where $\vec{t}_u$ contains the unreliable (missing) and $\vec{t}_r$ contains the reliable components of $\vec{t}$. $\vec{W}_{m,u}$ and $\vec{W}_{m,r}$ contain the rows of $\vec{W}_m$ associated with $\vec{t}_u$ and $\vec{t}_r$. Similarly, $\vec{\mu}_{m,u}$ and $\vec{\mu}_{m,r}$ denote the components of $\vec{\mu}_m$ corresponding to the missing and reliable components. The vector $\vec{x}$ includes representations of the signal in the

q-dimensional latent space. The MPPCA model is a mixture, such that each observation is fully accounted for by one unique element of that mixture. This means that it is necessary to determine the sub-model that best represents a given observation $\vec{t}$. Therefore, computing the optimal reconstruction boils down to finding the MAP estimate of the parameter set $\vec{\gamma} = \{\vec{S}, \vec{x}, \vec{t}_u\}$. This estimate can be obtained by solving:

$$\hat{\vec{\gamma}} = \underset{\vec{\gamma}}{\operatorname{argmax}} \left\{ P(\vec{\gamma}|\vec{t}_r) \right\} \tag{2.3}$$

This estimation problem is very similar to what we had in the training phase. Therefore, we might use Gibbs sampling to obtain the optimal values. However, for our proposed method, we have developed a a computationally efficient closed-form procedure for obtaining MAP estimates of the unknown parameters, based on expanding their posterior probabilities. We expand Eq. (2.3) to find the closed form solution; the detailed mathematics is presented in Appendix A. Based on the closed form expansions we arrive at the following procedure for recovering the missing features:

1. Using given $\vec{d}_r$ reliable components of an unknown signal $\vec{t}$, the Mahalanobis distances between the observed values and their counterparts in the mean vector of each sub-model are computed using

$$d_m = (\tau/2) \left(\vec{t}_r - \vec{\boldsymbol{\mu}}_{m,r}\right)^T \boldsymbol{\Lambda}_m \left(\vec{t}_r - \vec{\boldsymbol{\mu}}_{m,r}\right) \tag{2.4}$$

   with

$$\vec{\Lambda_m} = (\vec{I} - \tau \vec{W}_{m,r} \vec{\Sigma}_m \vec{W}_{m,r}^T) \tag{2.5}$$

   and

$$\vec{\Sigma}_m^{-1} = \vec{I}_q + \tau \vec{W}_{m,r}^T \vec{W}_{m,r}, \tag{2.6}$$

2. Combine the distances computed in step 1 with the prior probabilities of the sub-models to determine the most probable sub-model by

$$\hat{m} = \underset{m}{\operatorname{argmax}} \left\{ \log(\pi_m) - d_m \right\} \tag{2.7}$$

3. Having $\hat{m}$, estimate $\hat{\vec{x}}$ using

$$\hat{\vec{x}} = \vec{\eta}_{\hat{m}} = \tau \vec{\Sigma}_{\hat{m}} \vec{W}^T_{\hat{m},r} (\vec{t}_r - \vec{\mu}_{\hat{m},r}) \tag{2.8}$$

with

$$\vec{\Sigma}^{-1}_{\hat{m}} = \vec{I}_q + \tau \vec{W}^T_{\hat{m},r} \vec{W}_{\hat{m},r} \tag{2.9}$$

4. Recover the missing part of $\vec{t}$ by substituting $\hat{m}$ and $\hat{\vec{x}}$ in

$$\hat{\vec{t}}_u = \vec{W}_{\hat{m},u} \hat{\vec{x}} + \vec{\mu}_{\hat{m},u}. \tag{2.10}$$

5. If the estimated value of a component in $\hat{\vec{t}}_u$ is higher than the observed noisy value, the estimate is considered as incorrect. We use the observed values of $\vec{t}_u$ as the upper bound for imputation.

6. If there is no reliable feature in the input vector, the reconstructed vector is equal to the mean μ_i of the of the sub-model with the largest prior probability π_i. The sub-model with the highest prior probability mainly represents silence and very low energy speech.

The complete missing feature imputation procedure is summarized in Figure 2.2.

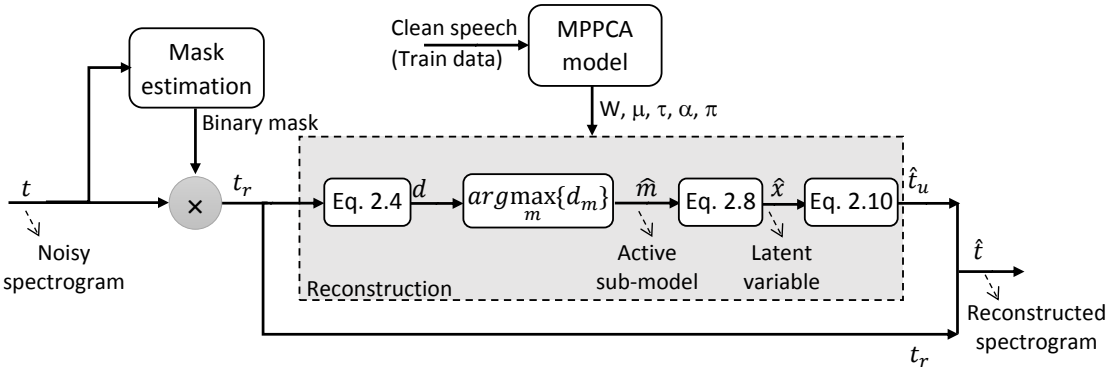


FIGURE 2.2: Missing data reconstruction using the MPPCA model

2.3 Experiments

To evaluate the performance of our proposed missing data imputation approach, the recognition performance on the MPPCA imputed spectrograms must be obtained. In this section we first present a brief overview of the mask estimation

procedure and the way in which we define the observation vectors used in the missing feature imputation procedure. Then, we summarize the commonalities and differences of the end-to-end speech recognition systems that were used in small vocabulary (AURORA-2) and large vocabulary (AURORA-4) tasks. Finally, we devote two sub-sections to these tasks.

2.3.1 Experimental setup

2.3.1.1 Mask estimation

An important aspect of all missing-feature approaches is the estimation of a spectrographic mask to distinguish the reliable components from unreliable ones. Various mask estimation techniques have been proposed in the literature (cf. section VI.D of Li et al. (2014)). In this research we used the estimate of the local SNR in individual time-frequency cells of the spectrogram. For estimating the noise power we used the MCRA2 method proposed by (Rangachari and Loizou, 2006). MCRA2 is designed for handling highly non-stationary environments. The probability that speech is present in a spectro-temporal cell is determined using the ratio between the momentary power of the signal and an estimate of the minimum power in that cell, which is updated continuously by averaging past values of the noisy speech power spectrum with a look-ahead factor. The SNR estimates appeared to be unreliable in the low-energy portions of the utterances. Therefore, we computed the logical AND of the MCRA2-based mask and the output of the voice activity detector (VAD) proposed in Ramírez et al. (2004).

Because in AURORA-2 and AURORA-4, used in this research, the clean speech versions of the noisified utterances are available, it is possible to compute the spectrogram of the noise. This allows us to obtain the so-called *oracle mask* by comparing the difference between the power of the clean speech and of the noise for all spectro-temporal cells to some suitably chosen threshold. The difference between the recognition accuracy obtained with the oracle mask and with the estimated mask is a powerful indicator of the efficacy of a missing feature imputation method.

2.3.1.2 Forming the observation vectors

Missing feature imputation procedures operate on spectro-temporal representations of speech signals. The MPPCA procedure operates on log-Mel-power spectra with 23 bands, obtained from 25 ms frames, shifted with 10 ms steps. Before the computation of the FFT, the frames are pre-emphasized and multiplied by a Hamming window. Speech spectra are characterized by a larger degree of continuity over time than most noise types. Therefore, the input for missing feature imputation methods are formed by stacks of consecutive 10 ms frames. The optimal number of consecutive time frames, indicated as the block size K, that span the observation space, as well as the number of frames over which the blocks are shifted, are not known in advance. In all experiments we fix the number of frames with which blocks are shifted to three frames, which corresponds to 30 ms. The optimal block size K is one of the parameters that we investigate in the experiments. With a step size of 3 frames for shifting blocks of feature frames, the minimum block size K is 3 frames. We will investigate block sized up to 30 frames, corresponding to the maximum block size used in Gemmeke et al. (2011b).

With a step size of 3 frames for shifting blocks and block size $K > 3$ frames, all frames –except the first two and last two frames of an utterance, which very likely correspond to silence– will be part of more than one block, and therefore subject to more than one imputation operation. The eventual reconstruction is formed by averaging the individual reconstructions. It is not necessarily true that a larger number of estimates in an average will shrink the confidence interval of the eventual estimate. The larger the block size, the larger, and as a consequence the more variable, the acoustic-phonetic context becomes. The detrimental effect of the variance due to the context may eliminate the beneficial effect of a larger number of observations. We will investigate the impact of K on the recognition accuracy.

2.3.1.3 System overview

ASR is the task of finding the most likely sequence of words, given some speech signal. Using the well-known Bayesian inversion, actual back-ends search for the sequence of words that maximize the likelihood of observing the speech signal. At a finer level of detail, words are modeled as sequences of speech sounds, so the

actual search is for the sequence of sounds that make up words and maximize the likelihood of the speech signal. The search is implemented by means of a finite state acceptor.

An overview of our experimental system is depicted in Figure 2.3. The log-Mel-spectra of the noisy speech are first processed using the MPPCA-based imputation. This process is summarized in the left-hand block of the diagram in Figure 2.3. The resulting 'cleaned' log-Mel-spectra form the input for the recognizer back-end in the right-hand block in that diagram. We use a GMM-HMM back-end for the AURORA-2 task. The digit words are modeled as sequences of 16 states that must be traversed from left to right. States cannot be skipped. Silence is modeled as a three-state HMM. The acoustic observations in the states are modeled by a mixture of six Gaussian distributions. Utterances are modeled as transitions from one digit word to the next. The AURORA-2 corpus was designed such that all digits and all digit-to-digit transitions occur approximately equally often. Recognition amounts to searching for the sequence of states that is most likely given the acoustic observations. For training the models of the individual states, the log-Mel-spectra are converted into Mel-frequency cepstral coefficients (MFCC). The GMM-HMM back-end is trained using the clean speech in the training corpus.

For the AURORA-4 task we use two back-ends. The first one uses GMM-HMMs trained with MFCC features for the sub-word units in combination with a finite state acceptor that implements a trigram language model. Here, the sub-word units are derived from a set of 39 phone symbols that were used in the phonetic transcription of the speech. The second back-end uses deep neural networks (DNN) for estimating the posterior probabilities of the sub-word units.[1] The DNNs are trained with log-Mel-spectrum features. The posterior probabilities computed by the DNNs are then used with the same finite state acceptor. The implementation details of the ASR back-ends are explained in Sections 2.3.2 and 2.3.3.

2.3.2 Digit string recognition on AURORA-2

We first applied the MPPCA imputation method to the AURORA-2 task with the default GMM-HMM recognition system described in Hirsch and Pearce (2000). The data set consists of connected digit utterances with lengths of one to seven

[1] The amount of training data in AURORA-2 is too small for training effective DNNs.

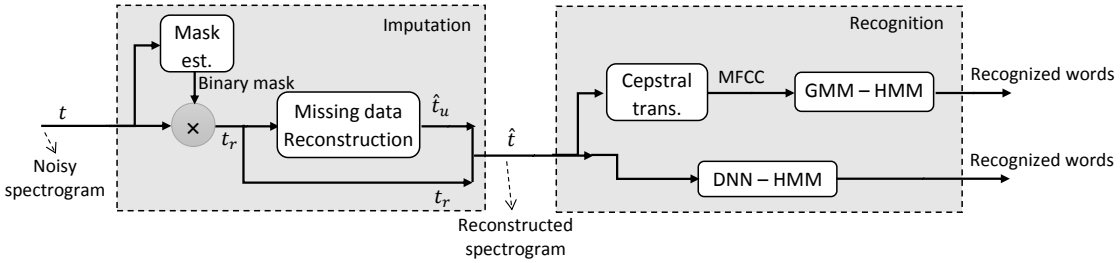


FIGURE 2.3: Diagram of the complete ASR system. Missing data imputation in combination with a recognition back-end. The GMM-HMM back-end is used in both experiments. The DNN-HMM back-end is only applied in the large vocabulary task.

digits per utterance. There are clean and multi-condition training sets available. The multi-condition training set contains the clean training data, plus the same signals noisified by four different noise types at SNR levels of $20, 15, 10, 5$ dB. The clean training set consisting of $8,440$ utterances was used for training the MPPCA models and the GMMs and transition probabilities in the recognizer back-end. The multi-condition set was used to determine the block size (K), the number of MPPCA sub-models (M), and the dimension of the manifold representations (q). We also used the multi-condition train set to tune the mask thresholds. We report results for test sets A and B before and after the application of MPPCA-based imputation. Both test sets contain 4004 utterances, in the clean version and noisified by four different noise type at SNR levels $-5, 0, 5, 10, 15, 20$ dB. The noise types in set A are the same as in the multi-condition training data; the noise types in set B are not represented in the training corpus.

For training and testing the GMM-HMM the 23-band log-Mel-spectra were converted to Mel-frequency cepstral coefficients (MFCC). MPPCA-based imputation was used to reconstruct log-Mel-spectra that contained missing features before the conversion to MFCCs. Figure 2.4 contains an example of the impact of MPPCA-based imputation on the log-Mel-spectrogram of a noisified utterance 'eight'. The SNR is 10 dB; the noise is suburban train noise. Panel (a) shows the spectrogram of the clean speech; panel (b) contains the spectrogram of the noisified signal. Panel (c) only contains the reliable spectro-temporal features, and panel (d) shows the MPPCA-based reconstruction. The application-dependent parameters of the MP-PCA model used for the imputation had the values $\{M = 30, K = 5, q = 20\}$. While the reconstructed spectrogram is clearly different from the clean spectrogram, it can be seen that the (visually) most salient features have been recovered. As in

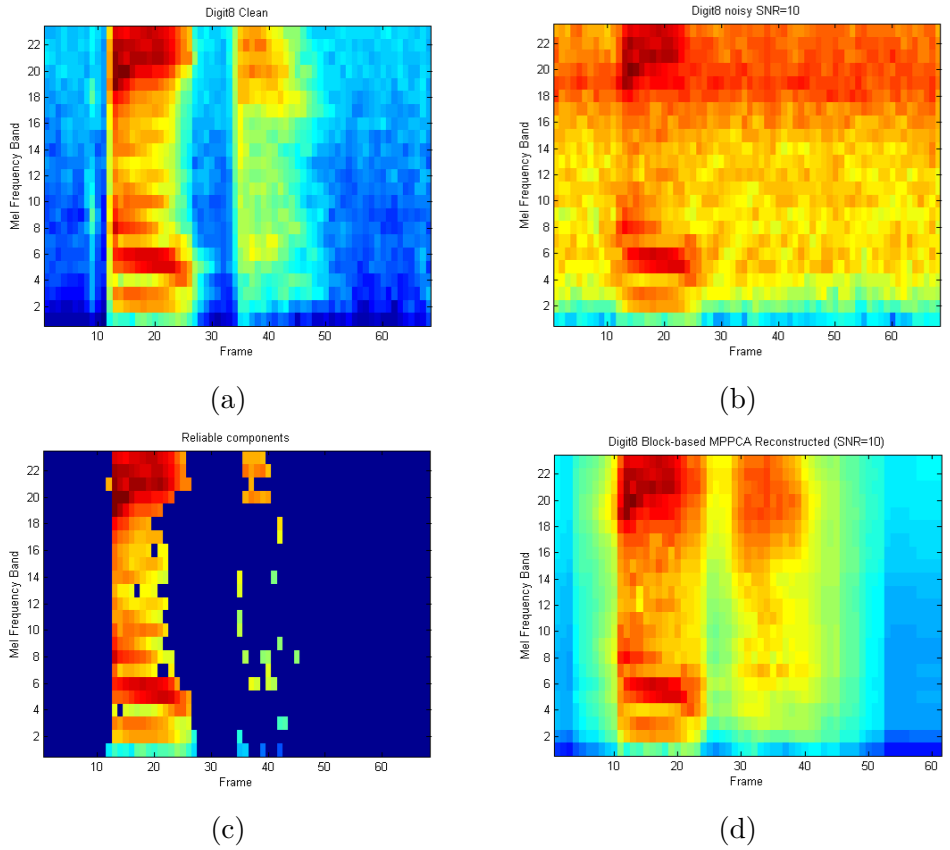


FIGURE 2.4: An example of a MPPCA imputed short noisy utterance. (a): the original clean spectrogram, (b): the spectrogram of the noisy utterance at SNR=10dB, (c): the reliable components, (d): the MPPCA imputed spectrogram. Imputation was performed with a MPPCA model with $\{K = 5, M = 30, q = 20\}$

all lossy coding, successful imputation yields spectrograms that tend towards the prototypical spectrogram of the digit words.

First, we determined a good value for the dimension of the PPCA sub-models. For that purpose we trained MPPCA models with $q = d$, $M = 11$ (motivated by the fact that there are eleven digit words), and $K = 5, 10, \cdots, 30$. We observed a sharp decrease of the values of $1/\alpha$ around $q = 20$. Subsequent recognition experiments confirmed that $q = 20$ suffices to represent the relevant variation in the acoustic space. With $q = 20$ we then performed a grid search over MPPCA models with $\{K, M\}$ pairs, with $K = 5, 10, \cdots, 30$ and $M = 1, 15, \cdots, 40$ on the multi-condition training data. The maximum block size $K = 30$ was imposed by the duration of the shortest utterance. The block shift was always fixed at 3 frames.

Unsurprisingly, there was no $\{M, K\}$ pair that yielded maximum recognition

accuracy on all 5 (SNR levels) × 4 (noise types) in the multi-condition training data. It appeared that accuracy reached a shallow plateau for values of M around 30, for oracle and estimated masks. Therefore, we only report results for M = 30. With the oracle mask the recognition accuracy decreased monotonically with increasing block size K. However, the reverse was true for the estimated mask: recognition accuracy kept increasing with increasing block size.

Table 2.1 summarizes the recognition results on test set A and test set B using a MPPCA model with q = 20 and M = 30, with block sizes K = 5 and K = 30. Intermediate values of K did not yield additional insights. The column labeled 'Baseline' contains the recognition results without any form of imputation or feature enhancement. The results are shown for both the oracle and the estimated masks.

	Test set A				
SNR	Baseline	5 Frames		30 Frames	
		Oracle	Estimated	Oracle	Estimated
20	94.09 (±0.41)	97.78 (±0.25)	90.24 (±0.51)	97.16 (±0.28)	95.06 (±0.37)
15	85.37 (±0.61)	97.30 (±0.28)	86.66 (±0.59)	95.85 (±0.34)	91.76 (±0.48)
10	65.35 (±0.83)	95.66 (±0.35)	78.20 (±0.72)	91.46 (±0.48)	81.40 (±0.67)
5	36.96 (±0.84)	91.46 (±0.48)	60.03 (±0.85)	80.01 (±0.69)	58.14 (±0.86)
0	14.39 (±0.61)	84.51 (±0.63)	34.96 (±0.83)	56.27 (±0.86)	29.38 (±0.79)
-5	7.52 (±0.46)	76.14 (±0.74)	15.97 (±0.63)	33.91 (±0.82)	13.08 (±0.58)
Average	50.61 (±0.35)	90.48 (±0.20)	61.01 (±0.34)	75.78 (±0.30)	61.47 (±0.34)
	Test set B				
20	94.25 (±0.40)	98.65 (±0.20)	91.96 (±0.47)	98.19 (±0.23)	96.64 (±0.31)
15	84.32 (±0.63)	98.33 (±0.22)	88.91 (±0.54)	97.32 (±0.28)	93.98 (±0.41)
10	61.22 (±0.85)	97.88 (±0.25)	80.65 (±0.68)	94.69 (±0.39)	85.08 (±0.62)
5	32.45 (±0.81)	95.87 (±0.34)	63.01 (±0.84)	85.74 (±0.61)	63.37 (±0.84)
0	13.02 (±0.58)	91.80 (±0.47)	36.22 (±0.83)	64.02 (±0.83)	35.16 (±0.83)
-5	7.27 (±0.45)	85.72 (±0.61)	16.23 (±0.64)	39.14 (±0.85)	15.45 (±0.63)
Average	48.75 (±0.35)	94.71 (±0.16)	62.83 (±0.34)	79.85 (±0.28)	64.94 (±0.34)

TABLE 2.1: Average ASR accuracies for MPPCA imputation on AURORA-2. 95% confidence intervals in parentheses.

2.3.2.1 Interpretation

Thanks to the fact that the MPPCA method is noise ignorant, the recognition accuracy for test sets A and B are similar. In a previous application of CS to noise-robust ASR (Gemmeke et al., 2011b), the performance for test set B was

clearly inferior. If anything, in our data the accuracy in test set B is higher than in test set A. The only parameter that could be (and indeed is) sensitive to the differences between the noise types is the threshold used in determining whether a spectro-temporal measurement is reliable.

From Table 2.1 it is clear that MPPCA-based imputation improves the recognition accuracy in all SNR conditions. Also, imputation with the oracle mask always outperforms the estimated mask. The relative loss with the estimated mask increases sharply with decreasing SNR level. The results with the oracle mask are always better for 5-frame blocks than for 30-frame blocks. The accuracies of 76.14% in test set A and 85.72% in test set B at $SNR = -5\,\mathrm{dB}$ with five-frame blocks suggest that the MPPCA imputation yields excellent reconstructions, even if the number of reliable features is very small. The oracle mask guarantees that there are no false positives. The performance with the oracle mask with the 30-frame blocks suggests that there is indeed an effect of context smearing. That effect is already visible at $SNR = 20\,\mathrm{dB}$. The effect becomes worse with decreasing SNR level. At the lowest SNR levels the number of reliable features is small. If the reliable features happen to be at a large distance from the frame under analysis, the average of the 30 estimates is determined biased towards the spectra of the remote frames. Even in the absence of false accepts it is difficult to reconstruct spectral vectors on the basis of reliable features scattered randomly over 30 by 23 pixel patches if these patches are sparsely filled.

For the estimated mask the impact of block size is more complex. Although the mask threshold was set to a level that minimizes false accepts –at the cost of a substantial proportion of false rejects, which are less harmful– it is not possible to completely prevent false accepts with the estimated mask. The relative proportion of false accepts increases with decreasing SNR. With the 5-frame blocks the effect of a decreasing number of reliable features, in combination with an increasing proportion of false accept as SNR decreases, is evident from the highest SNR levels. For $SNR > 5\,\mathrm{dB}$ 30-frame blocks outperform the 5-frame blocks. So much so that, averaged over the six SNR levels, the 30-frame blocks come out as the winner. Most probably, the availability of a larger number of correctly identified reliable features, along with a relatively small number of false accepts, overcomes the detrimental effect of context smearing. In the conditions with $SNR \leq 5\,\mathrm{dB}$ the smearing effect becomes dominant.

In summary, the experiment on AURORA-2 confirms the efficacy of MPPCA-based imputation. However, in conditions with $SNR \leq 5\,\text{dB}$ the effect of a growing proportion of false accepts in combination with a decreasing number of features that are considered as reliable is increasingly more difficult to overcome.

2.3.3 Large Vocabulary continuous speech

AURORA-4 (Parihar and Picone, 2002; Parihar et al., 2004) is a good platform for investigating noise robustness in large vocabulary continuous speech recognition. The database contains training and test sets that combine additive noise at several SNR levels with channel variation. All speech was recorded with a Sennheiser HMD-414 close-talking microphone and one of 18 different microphones in parallel. The data is sampled with 16 kHz sampling frequency. There are two training sets, each comprised of the same 7,138 utterances. We used train set 1 (identical to the WSJ0 SI-84 set) for training the MPPCA models and a GMM-HMM recognizer; train set 1 only includes clean data recorded with the Sennheiser microphone. Train set 2 – the multi-condition training set – contains the same 7,138 utterances. Half of the utterances were recorded with the Sennheiser microphone, the other half with one of 18 different microphones. One-fourth (2×893 utterances) of the Sennheiser and alternative microphone were clean speech. The remaining three-fourths ($2 \times 2,676$ utterances) were corrupted by adding six different noise types (car, babble, restaurant, street, airport, and train) at randomly selected SNRs between 10 and 20 dB. We used the multi-condition training set 2 for training a GMM-HMM multi-condition trained recognizer and a DNN-HMM recognizer.

The test set consist of 14 subsets, each containing 330 utterances. Subsets T-01 to T-07 were recorded with the Sennheiser microphone and T-08 to T-14 were recorded with another microphone. In each group one subset is clean and the remaining six subsets are artificially noisified with the six noise types at varying SNRs between 5 and 15 dB in steps of 1 dB. A development set of the same structure as the test set, but with different utterances, is available. We used this subset for optimizing the model parameters M, q and K.

For the AURORA-4 experiments, the 'Complete recipe' recognizers in the Kaldi toolkit (Povey et al., 2011) were used. That procedure includes the training of a trigram language model. The computation of reconstructed log-Mel spectrogram

and MFCC features was performed according to the procedure in Figure 2.3. Cepstral mean and variance normalization was applied per utterance.

2.3.3.1 GMM-HMM back-end

The HMM-GMM-based decoder for AURORA-4 makes use of context-dependent tied-state triphone models. Three-state models establish around 2000 distinct HMM states. The raw features consisted of 7 consecutive frames of 13 MFCCs, and the dimensionality of the features was reduced to 39 using linear discriminant analysis (LDA) (Haeb-Umbach and Ney, 1992). Next, speaker normalization is performed using the maximum-likelihood linear transform (MLLT) (Saon et al., 2000). We experimented with two different versions of the GMM-HMM recognizer. First, as is common in the missing data imputation field, we trained the recognizer with the clean training set. Evaluating the performance of this system on MPPCA imputed features shows to what extent the imputation has been successful in reconstructing the corrupted speech features. However, it is known that multi-condition training always outperforms imputation, especially in the lower SNR levels (Acero, 1993; Kolossa and Haeb-Umbach, 2011; Virtanen et al., 2012). Therefore, we decided to include a comparison of the accuracy obtained with a multi-condition trained recognizer that operates on raw and on imputed features. For the multi-condition training the raw features were used.

2.3.3.2 DNN-HMM back-end

In the DNN-HMM hybrid system, the posterior probability estimates for the HMM states are provided by the trained DNNs (Hinton et al., 2012). The Kaldi recognizer uses the nnet2 DNN implementation described in Zhang et al. (2014) with four hidden layers using p-norm units with 3,000 input and 300 output dimensions. The last (output) layer is a softmax layer with a dimension equal to the number of context-dependent states (2,000 HMM states in our case). We trained the DNN with the multi-condition training data. The training procedure uses an approach similar to greedy layer-wise supervised training (Bengio et al., 2007) or the layer-wise back-propagation (Seide et al., 2011). The network was randomly initialized with one hidden layer and trained for less than one epoch. Then the layer of weights that go to the softmax layer was removed. A new hidden layer

and two sets of randomly initialized weights were added and trained again. This was repeated until the desired number of layers was achieved (Zhang et al., 2014).

2.3.3.3 Training and Tuning of the MPPCA model for large vocabulary ASR

The basic features, i.e., stacks of a number of frames of 23-dimensional log-Mel spectra, do not differ between the AURORA-2 and the AURORA-4 task. Still, the dimensionality of the latent space and the number of sub-models that are needed for a good approximation of the manifold that supports the speech signals may differ between the tasks, because of the nature of the task (continuous speech versus connected words) and the much larger number of words in AURORA-4. Therefore, we did experiments with values of q and M that are larger than the best values obtained with AURORA-2. Since intermediate block sizes did not yield interesting results in AURORA-2, we only experimented with block sizes of 5 and 30 frames. In the search for optimal parameter values we only performed recognition on the development set with the GMM-HMM back-end. We assumed that MPPCA-based imputation would not affect the log-Mel-spectra to such an extent that the DNNs, which were trained with clean speech recorded with the Sennheiser microphone, was no longer able to yield appropriate posterior probability estimates. The optimal value of q again was 20; the optimal value of M was 35. The difference in recognition accuracy between $K = 5$ and $K = 30$ was negligible. Thus, all results in Table 2.2 are for $\{q = 20, M = 35, K = 5\}$.

It is interesting to see that the dimension q of the PPCA sub-models did not change when going from a connected digits to a large vocabulary task. Also the number of PPCA sub-models that are needed for an adequate approximation of the speech manifold does not increase substantially for large vocabulary speech. Apparently, a much larger amount of variability in the phonetic context does not give rise to a much more complex structure of the manifold.

2.3.3.4 Results and discussion

The results of the experiments with AURORA-4 are summarized in Table 2.2. To avoid cluttering up an already very busy table, we present the 95% confidence intervals for a number of relevant accuracy values in Table 2.3. The number of

words in the individual test sets is about 5,300. The total number of words –the basis for calculating the confidence interval of the grand average– is 74,942. From Table 2.2 it can be seen that MPPCA imputation is always beneficial, even if the back-end is trained with the multi-condition data. This holds both for the GMM and the DNN systems. Also, MPPCA-based imputation appears to be beneficial in the clean conditions (test sets T-01 and T-08). We conjecture that this is because MPPCA-based reconstruction causes some degree of normalization of the feature vectors. The results with clean training are worse in test sets T-08 – T-14 than in test sets T-01 – T-07 because we only used the clean signals recorded with the Sennheiser microphone for training. The difference disappears in the multi-condition training, where the alternative microphones are represented in the training data. From a comparison of the accuracies obtained with the baseline system and the MPPCA-based system that used the oracle mask it can be seen that imputation improves accuracy for test sets T-08 – T14 to the same extent as test sets T01 – T07.

From the comparison between the GMM-HMM and DNN-HMM results in Table 2.2 it can be seen that the gain from using DNNs instead of GMMs with multi-condition training is small. The largest gains are obtained in test sets T-09 – T-14. This suggests that the DNNs are better able to cope with the additional variation introduced by the alternative microphones. It is also clear that MPPCA-based imputation is always beneficial in the DNN-based system.

In Table 2.2 it can be seen that –as expected– the recognition accuracy with the estimated mask is always below the performance with the oracle mask. At the end of Section 2.1 it was stated that the loss when replacing the oracle mask by the estimated mask is an important evaluation criterion for the efficacy of an MDT approach. To assess the efficacy of the proposed MPPCA-based imputation it is useful to compare the loss of accuracy due to mask estimation errors with a recently published alternative imputation technique that uses a truncated Gaussian mixture model instead of a mixture of probabilistic PCAs (González et al., 2012). In Table 2.4 the accuracies reported for the 'Imputation' system of González et al. (2012) are shown for their oracle and estimated masks (rows labeled 'Imputation' in their Table 2). Direct comparisons of the absolute accuracies appears to be very difficult, if only because the mask estimation procedure and the back-ends differ. Already the accuracies reported for their baseline system (cf. the row labeled 'baseline' in Table 2 in González et al. (2012)) differ from

	GMM-HMM						DNN-HMM		
	Clean trained			Multi-condition trained			Multi-condition trained		
	Baseline	MPPCA		Baseline	MPPCA		Baseline	MPPCA	
		Oracle	Est.		Oracle	Est.		Oracle	Est.
T-01	93.41	94.17	93.74	93.74	94.36	94.15	93.84	95.76	95.91
T-02	80.05	93.13	85.65	94.73	95.22	94.97	93.31	95.24	95.05
T-03	54.92	83.64	62.30	89.54	93.69	90.83	89.65	94.79	92.19
T-04	34.21	71.68	49.50	85.45	92.88	85.95	86.36	93.95	89.58
T-05	49.41	77.92	56.66	86.49	93.37	87.09	87.54	94.17	90.21
T-06	60.92	87.78	68.32	90.58	93.65	92.00	90.92	94.84	92.99
T-07	54.08	78.85	60.88	87.46	92.92	88.06	88.38	94.21	90.45
T-08	56.85	58.42	57.67	88.01	87.75	87.75	89.16	92.45	92.86
T-09	55.65	64.88	61.80	83.02	85.45	83.50	86.42	89.43	87.82
T-10	40.31	57.71	47.56	76.69	84.42	78.48	78.37	87.33	81.58
T-11	26.45	47.23	36.73	74.80	83.90	75.17	76.46	84.33	78.40
T-12	35.72	52.21	41.70	72.31	82.83	74.50	77.04	85.67	79.47
T-13	42.46	61.16	50.03	77.96	84.40	78.57	78.65	87.37	81.82
T-14	37.59	54.10	44.80	73.79	81.52	74.69	76.65	85.00	78.87
Avg.	51.57	70.20	58.38	83.90	89.02	84.69	85.20	91.04	87.66

TABLE 2.2: ASR accuracies for MPPCA imputation (5-Frames) on the AURORA-4 test sets.

Accuracy range (%)	35	45	55	65	75	85	95
95% Confidence interval (T 01–14)	1.30	1.35	1.35	1.30	1.18	0.97	0.59
95% Confidence interval (Avg)	0.34	0.36	0.36	0.34	0.31	0.26	0.16

TABLE 2.3: The 95% confidence interval for a few ranges of accuracies both for test subsets (5,353 words) and the average (74,942 words).

our baseline system. For the clean test set T-01 (Sennheiser microphone) we obtained an accuracy of 93.41%, which compares favorably with 87.26% in González et al. (2012). For the clean test set T-08 (alternative microphones) we obtain an accuracy of 56.85%, compared to 76.42% in González et al. (2012). From these data it can be inferred that González et al. (2012) also used the clean data from the alternative microphones in train set 2 for training their back-end, while our system only used the Sennheiser microphone recordings in train set 1. Including recordings made with the alternative microphones has a negative effect in test set T-01, but it improves the accuracies in test sets T-08 - T-14. Different training material might also explain the other performance differences between their and our oracle system for the test sets T-02 – T-07. To further illustrate how difficult it is to compare published systems we refer to a follow-up paper from the same group (Gonzalez et al., 2013), in which they present slightly improved results using a similar truncated Gaussian approach (the TGI system). In that paper they show substantially improved baseline results.

Set	González et al. (2012)				MPPCA	
	Oracle	Estimated	diff	ratio	diff	ratio
T-01	87.26	87.00	0.26	0.99	0.43	0.99
T-02	85.45	55.86	29.59	0.65	7.48	0.93
T-03	84.53	58.47	26.06	0.69	21.34	0.74
T-04	86.31	80.93	5.38	0.93	22.18	0.69
T-05	84.10	52.98	31.12	0.63	21.26	0.73
T-06	83.71	59.07	24.64	0.70	19.46	0.78
T-07	83.00	61.70	21.30	0.74	17.97	0.77
T-08	76.42	73.01	3.41	0.95	0.75	0.99
T-09	73.68	46.65	27.03	0.63	3.08	0.95
T-10	73.64	50.05	23.59	0.67	10.15	0.82
T-11	76.26	67.79	8.47	0.88	10.50	0.78
T-12	72.65	45.45	27.20	0.62	10.51	0.80
T-13	70.60	48.09	22.51	0.68	11.13	0.82
T-14	70.99	53.37	17.62	0.75	9.30	0.83
Average	79.18	60.03	19.15	0.75	11.82	0.83

TABLE 2.4: Comparison of differences between the accuracy obtained with oracle and estimated masks on the AURORA-4 task. The columns 'Oracle' and 'Estimated' contain the accuracies obtained with the 'Imputation' system in González et al. (2012). The columns 'diff' (oracle - estimated) and 'ratio' (estimated/oracle) show the deterioration in the 'Imputation' (left) and the MPPCA system (right). The numbers corresponding to MPPCA system are computed from 2nd and 3rd columns of Table 2.2. (The recognizer back-ends are clean trained GMM-HMM)

To compensate for the fact that the oracle accuracies differ between our system and the 'Imputation' system of González et al. (2012), we computed two indicators of the deterioration between the oracle and the estimated masks: the absolute difference in accuracy, and the ratio between the accuracy with the estimated mask and the oracle mask. Higher values for the ratio indicate less deterioration. As can be seen in Table 2.4, both indicators show the same trend. It can be seen that in the MPPCA-based system the average absolute loss is smaller. This is to a large extent due to larger losses in test sets T-08 – T-14 in González et al. (2012), which in turn are due to very high accuracies with the oracle mask. Therefore, it is safe to assume that the MPPCA imputation method does not suffer from the fact that the alternative microphones were not represented in the training material. However, the mismatch between training on speech recorded with the Sennheiser microphone and testing on speech recorded with multiple different microphones does affect the GMM-based recognizer back-end. Interestingly, the same conclusions hold for the comparison of the TGI system in Gonzalez et al. (2013) and our MPPCA-based system.

In summary, the data in Table 2.4 show that the loss between oracle and estimated mask with MPPCA-based imputation is almost always smaller than the imputation results in González et al. (2012) and Gonzalez et al. (2013).

2.4 General Discussion

In this chapter we proposed a novel manifold-based compressive-sensing approach to missing data imputation, in which we used a mixture of linear probabilistic principal component analyzers to approximate the unknown non-linear manifold that supports the (undistorted) speech signals. The original observation space, i.e., segments of spectrograms, is treated as the embedding space. This is different from previous applications of compressive sensing to speech processing that *generated* an embedding space from the observation data. To the best of our knowledge, our work is the first attempt to use MPPCA to learn the low-dimensional non-linear manifold that supports spectrographic representations of speech signals.

To be able to train an MPPCA model, two application-dependent parameters must be fixed, viz. the number of consecutive speech frames that form a feature vector in the observation space (K) and the number of probabilistic PCAs that form the mixture that approximates the nonlinear manifold (M). A third parameter, the dimension of the PPCA models (q), can be inferred from the values in the parameter vector $\vec{\alpha}$ in Figure 2.1. It appeared that a dimension $q = 20$ for the latent space is sufficient, both for the AURORA-2 and the AURORA-4 tasks. We also found that the number of sub-models needed for approximating the manifold did not differ much between the 11-word vocabulary AURORA-2 and the 5000-word AURORA-4 tasks. The eleven digit words in AURORA-2 contain only 19 of the 39 basic symbols that are needed for transcribing spoken American English.

The finding that AURORA-2 and AURORA-4 both require about 30 PPCAs to approximate the manifold suggests that there is no straightforward relation between the number of phones and the number of sub-models. A mixture of $30 \sim 35$ PPCA analyzers yields a much more compact representation than the 231 surface-level head-body-tail models in older attempts to capture the acoustic structure of connected digit utterances for improving recognition performance (Gandhi and Jacob, 1998; Han et al., 2007). It also compares favorably against a number of 256

Gaussians in GMM-based missing feature imputation (González et al., 2012; Gonzalez et al., 2013). Basically, the MPPCA model is also a mixture of Gaussians, but it differs from conventional Gaussian mixtures because of the particular structure imposed on the distributions. Their covariance is effectively low rank and the rank is equal to the intrinsic dimensions of the manifold. This property makes our method capable of finding the position of data points on the manifold using a small number of reliable components (of the order of intrinsic dimensionality).

Although our results with respect to the block size (K) are somewhat equivocal, it is interesting to observe that in most conditions excellent results were obtained with 5-frame blocks. Gonzalez et al. (2013) also found that a 5-frame window yielded the best results in their GMM-based imputation procedure. Blocks of five consecutive frames cover a time window of about 50 ms, which is slightly shorter than the average duration of phones in continuous speech. In the compressive sensing approach proposed in Gemmeke et al. (2011b) the best results were obtained with much longer windows that spanned up to 300 ms. The long windows were needed to distinguish the long-term continuity of the acoustic features of speech signals from the usually much shorter continuous feature traces in background noise. For MPPCA (and TGI) much shorter windows suffice, because the focus is on the (articulatory-induced) acoustic patterns in clean speech that are mainly determined by transitions between neighboring phones, independent of corrupting background noise.

In the taxonomy proposed in Li et al. (2014) the MPPCA method belongs in the category of approaches that do not use prior knowledge about the corrupting noise. Still, from a detailed analysis of the recognition results it appeared that the optimal mask threshold with the MCRA2 mask differs somewhat between the noise types. In future research we plan to experiment with different thresholds in a number of frequency bands. Our accuracy scores on AURORA-2 confirm the finding in previous experiments with missing feature imputation that there is an SNR level below which imputation methods that do not take into account prior knowledge about the noise lose out against competing methods that do take knowledge about the noise into account (e.g. Gemmeke et al., 2011b; Gonzalez et al., 2013). At SNR levels $< 0\,\mathrm{dB}$ the number of reliable features becomes very small. Yet, in Table 2.1 it can be seen that the recognition accuracies with the oracle mask are pretty high. Therefore, it can be concluded that MPPCA-based imputation yields excellent results in the absence of mask errors. At the same time it is clear that

the recognition accuracy drops to very low levels due to the inevitable increase of the proportion of false accepts in the lowest SNR conditions.

A direct comparison between the performance of MPPCA-based imputation in AURORA-2 and AURORA-4 is very difficult, because the lowest SNR condition in AURORA-4 is 5 dB, i.e., the level above which there appear to be sufficient reliable features to enable effective imputation. The impact of mask estimation errors, and the imputation errors this may cause differs very much between the two tasks. Averaged over the SNR levels from 20 dB to 5 dB, i.e., the SNR levels that the two tasks have in common, the ratio of the accuracy with the estimated mask and the accuracy with the oracle mask is 0.84 for AURORA-2 and 0.80 for AURORA-4. That the ratio is somewhat smaller in AURORA-4 is most likely due to the fact that the perplexity of the search in AURORA-4 is larger, which will make the search more sensitive to the effects of imputation errors.

Because mask estimation deeply affects all missing feature imputation approaches, improving mask estimation or mitigating the effects of estimation errors is essential for practical applications. We have already mentioned the possibility of using different mask thresholds in different frequency bands. Another way for dealing with mask estimation errors is replacing binary decisions about the reliability of spectro-temporal measurements by probabilistic estimates of the reliability. It is possible to extend the procedure for the reconstruction of clean spectrograms in the MPPCA method in such a way that it uses all features in an observation vector, but weighted with an estimate of their reliability instead of weights that are either zero or one. A similar procedure has proved to be beneficial in González et al. (2012); Gonzalez et al. (2013). Also, the probabilistic formulation of the MPPCA approach makes it possible to integrate the mask estimation with the reconstruction procedure. This could be done by calculating the likelihood of individual observation components given the trained model and considering the components with higher likelihood as the reliable components of the spectrogram.

Comparing the impact of mask estimation errors between competing missing data imputation methods appears to be next to impossible if these imputation methods are part of different ASR systems. It is impossible to decide whether observed differences are due to the imputation method or to different mask errors. Moreover, these comparisons do not advance our understanding of the way in which mask estimation errors interfere with recognition performance. Therefore,

it would be a large step forward if a method could be devised that deteriorates a given oracle mask in a known and controllable manner. Essential parameters to control are the proportion of mask errors, and the distribution of the errors over the time-frequency plane. An important aspect of that distribution is the trade-off between concentration in contiguous time-frequency regions versus the degree of scatter. It goes without saying that artificially induced mask estimation errors should be restricted to the areas in the time-frequency plane where the speech power is relatively small.

The research in this chapter was based on the assumption that our novel MPPCA-based imputation can reconstruct a noise-corrupted spectrogram that is sufficiently accurate to treat the output as clean. This, in turn, allows us to use the imputation result as input for a back-end that is trained with clean speech. It would be worthwhile to train the GMMs and the DNNs in the back-end with a mix of clean speech and noisy speech processed with MPPCA-based imputation. Because the MPPCA model is trained on clean speech only, and the imputation result is only mildly –if at all– dependent on the noise types, such an approach would avoid most of the limitations of multi-condition training.

The prior distributions imposed on some of the parameters of the MPPCA model (such as α) are selected to guarantee that the representations of the speech data in the sub-models is sparse. In the approach presented in this chapter some of the other prior distributions were chosen to be conjugate with the posterior distributions (for example the priors chosen for $\vec{\mu}$ and $\vec{\tau}$). Therefore, the prior distributions were selected independently of the characteristics of the signals of interest, i.e. speech signals. Future research will investigate whether prior knowledge of the distributions of the acoustic features of speech signals can be used in selecting the prior distributions imposed on the parameters of the MPPCA model. If that appears to be possible, the resulting models are likely to capture the important characteristics of speech signals better than models that ignore prior speech knowledge.

In this chapter manifold-based learning was only used for learning the latent space that formed the basis for missing feature imputation based on compressive sensing. Because the projection of the observed spectra onto the learned manifold induces a substantial degree of normalization, it is interesting to investigate the

possibility of using MPPCA as an alternative for existing approaches to speaker and channel normalization.

2.5 Conclusion

In this chapter we introduced a manifold-based compressive sensing approach to the imputation of missing features in noise-robust ASR. In this approach the observation space is considered as an over-complete embedding space of the signals on a lower-dimensional manifold in a latent space. The latent space is learned in the form of a mixture of probabilistic principal component analyzers. As long as there is a sufficient number of reliable measurements in the embedding space, the position of an observation (a slice of a spectrogram) can be located on the manifold. Compressive sensing theory then guarantees that it is possible to reconstruct the missing measurements in the embedding space, provided that the number of reliable features in the observation space is large enough. We developed a computationally efficient procedure for implementing the reconstruction.

Experiments on a connected digits recognition task (AURORA-2) and on a large vocabulary continuous speech recognition task (AURORA-4) showed that the MPPCA-based imputation using a mixture of $30 \ldots 35$ probabilistic principal component analyzers is at least as effective as competing GMM-based imputation methods that use a mixture of 256 Gaussians. MPPCA-based imputation seems to be less sensitive to inevitable mask estimation errors than imputation based on truncated Gaussians. The MPPCA approach can be improved by integrating it with a probabilistic mask estimation method.

Chapter 3

Sparse Coding of the Modulation Spectrum for Noise-Robust Automatic Speech Recognition

The full modulation spectrum is a high-dimensional representation of one-dimensional audio signals. Most previous research in automatic speech recognition converted this very rich representation into the equivalent of a sequence of short-time power spectra, mainly to simplify the computation of the posterior probability that a frame of an unknown speech signal is related to a specific state. In this chapter we use the raw output of a modulation spectrum analyser in combination with sparse coding as a means for obtaining state posterior probabilities. The modulation spectrum analyser uses 15 gammatone filters. The Hilbert envelope of the output of these filters is then processed by nine modulation frequency filters, with bandwidths up to 16 Hz. Experiments using the AURORA-2 task show that the novel approach is promising. We found that the representation of medium-term dynamics in the modulation spectrum analyser must be improved. We also found that we should move towards sparse classification, by modifying the cost function in sparse coding such that the class(es) represented by the exemplars weigh in, in addition to the accuracy with which unknown observations are reconstructed. This creates two challenges: (1) developing a method for dictionary learning that takes the class occupancy of exemplars into account and (2) developing a method for learning a mapping from exemplar activations to state

posterior probabilities that keeps the generalization to unseen conditions that is one of the strongest advantages of sparse coding.

3.1 Introduction

Nobody will seriously disagree with the statement that most of the information in acoustic signals is encoded in the way in which the signal properties change over time, while the instantaneous characteristics such as the shape or the envelope of the short-time spectrum, are less important - though surely not unimportant. The dynamic changes over time in the envelope of the short-time spectrum are captured in the modulation spectrum (Drullman et al., 1994; Hermansky, 1997; Xiao et al., 2008). This makes the modulation spectrum a fundamentally more informative representation of audio signals than a sequence of short-time spectra. Still, most approaches in speech technology, whether it is speech recognition, speech synthesis, speaker recognition, or speech coding, seem to rely on impoverished representations of the modulation spectrum in the form of a sequence of short-time spectra, possibly extended with explicit information about the dynamic changes in the form of delta and delta-delta coefficients. For speech (and audio) coding, the reliance on sequences of short-time spectra can be explained by the fact that many applications (first and foremost telephony) cannot tolerate delays in the order of 250 ms, while full use of modulation spectra might incur delays up to a second. What is more, coders can rely on the human auditory system to extract and utilize the dynamic changes that are still retained in the output of the coders. If coders are used in environments and applications in which delay is not an issue (music recording, broadcast transmission), we do see a more elaborate use of information linked to modulation spectra (Thompson and Atlas, 2003; Paliwal et al., 2011; Pichevar et al., 2011). Here too, the focus is on reducing bit rates by capitalizing on the properties of the human auditory system. We are not aware of approaches to speech synthesis - where delay is not an issue - that aim to harness advantages offered by the modulation spectrum. Information about the temporal dynamics of speech signal by means of shifted delta cepstra has proven beneficial for automatic language and speaker recognition (Torres-Carrasquillo et al., 2002).

In this chapter we are concerned with the use of modulation spectra for automatic speech recognition (ASR), specifically noise-robust speech recognition. In

this application domain, we cannot rely on the intervention of the human auditory system. On the contrary, it is now necessary to automatically extract the information encoded in the modulation spectrum that humans would use to understand the message.

The seminal research by Drullman et al. (1994) showed that modulation frequencies > 16 Hz contribute very little to speech intelligibility. In Arai et al. (1999) it was shown that attenuating modulation frequencies < 1 Hz does not affect intelligibility either. Very low modulation frequencies are related to stationary channel characteristics or stationary noise, rather than to the dynamically changing speech signal carried by the channel. The upper limit of the band with linguistically relevant modulation frequencies is related to the maximum speed with which the articulators can move. This insight gave rise to the introduction of RASTA filtering in Hermansky et al. (1991) and Hermansky and Morgan (1994). RASTA filtering is best conceived of as a form of post-processing applied on the output of otherwise conventional representations of the speech signal derived from short-time spectra. This puts RASTA filtering in the same category as, for example, Mel-frequency spectra and Mel-frequency cepstral coefficients: engineering approaches designed to efficiently approximate representations manifested in psycho-acoustic experiments (Hermansky, 2011). Subsequent developments towards harnessing the modulation spectrum in ASR have followed pretty much the same path, characterized by some form of additional processing applied to sequences of short-time spectral (or cepstral) features. Perhaps somewhat surprisingly, none of these developments have given rise to substantial improvements of recognition performance relative to other engineering tricks that do not take guidance from knowledge about the auditory system.

All existing ASR systems are characterized by an architecture that consists of a frontend and a back end. The back end always comes in the form of a state network, in which words are discrete units, made up of a directed graph of subword units (usually phones), each of which is in turn represented as a sequence of states. Recognizing an utterance amounts to searching the path in a finite-state machine that has the maximum likelihood, given an acoustic signal. The link between a continuous audio signal and the discrete state machine is established by converting the acoustic signal into a sequence of likelihoods that a short segment of the signal corresponds to one of the low-level states. The task of the frontend is to convert the signal into a sequence of state likelihood estimates, usually at a 100-Hz rate,

which should be more than adequate to capture the fastest possible articulation movements.

Speech coding or speech synthesis with a 100-Hz frame rate using short-time spectra yields perfectly intelligible and natural-sounding results. Therefore, it was only natural to assume that a sequence of short-time spectra at the same frame rate would be a good input representation for an ASR system. However, already in the early seventies, it was shown by Jean-Silvain Liénard (Mlouka and Liénard, 1974) that it was necessary to augment the static spectrum representation by so-called delta and delta-delta coefficients that represent the speed and acceleration of the change of the spectral envelope over time and that were popularized by Furui (1981). For reasonably clean speech, this approach appears to be adequate.

Under acoustically adverse conditions, the recognition performance of ASR systems degrades much more rapidly than human performance (Lippmann, 1996). Convolutional noise can be effectively handled by RASTA-like processing. Distortions due to reverberation have a direct impact on the modulation spectrum, and they also cause substantial difficulties for human listeners (Houtgast and Steeneken, 1985; Rennies et al., 2011). Therefore, much research in noise-robust ASR has focused on speech recognition in additive noise. Speech recognition in noise basically must solve two problems simultaneously: (1) one needs to determine which acoustic properties of the signal belong to the target speech and which are due to the background noise (the source separation problem), and (2) those parts of the acoustic representations of the speech signal which are not entirely obscured by the noise must be processed to decode the linguistic message (speech decoding problem).

For a recent review of the range of approaches that has been taken towards noise-robust ASR, we refer to Virtanen et al. (2012). Here, we focus on one set of approaches, guided by the finding that humans have less trouble recognizing speech in noise, which seems to suggest that humans are either better in source separation or in latching on to the speech information that is not obscured by the noise (or in both). This suggests that there is something in the auditory processing system that makes it possible to deal with additive noise. Indeed, it has been suggested that replacing the conventional short-time spectral analysis based on the fast Fourier transform by the output of a principled auditory model should improve robustness against noise. However, up to now, the results of research along

this line have failed to live up to the promise (Ghitza, 1994). We believe that this is at least in part caused by the fact that in previous research, the output of an auditory model was converted to the equivalent of the energy in one-third octave filters, necessary for interfacing with a conventional ASR back end, but without trying to capture the continuity constraints imposed by the articulatory system. In this conversion most of the additional information carried by the modulation spectrum is lost.

In this chapter we explore the use of a modulation spectrum frontend that is based on time-domain filtering that does not require collapsing the output to the equivalent of one-third octave filters, but which still makes it possible to estimate the posterior probability of the states in a finite-state machine. In brief, we first filter the speech signal with 15 gammatone filters (roughly equivalent to one-third octave filters) and we process the Hilbert envelope of the output of the gammatone filters with nine modulation spectrum filters (Jørgensen and Dau, 2011). The 135-dimensional (135-D) output of this system can be sampled at any rate that is an integer fraction of the sampling frequency of the input speech signal. We refer to the resulting vectors as envelope modulation spectrum (EMS) features. For the conversion of the EMS feature vectors to posterior probability estimates of a set of states, we use the sparse coding (SC) approach proposed by Gemmeke et al. (2011b). Sparse coding is best conceived of as an exemplar-based approach (Hurmalainen et al., 2011a) in which unknown inputs are coded as positive (weighted) sums of items in an exemplar dictionary.

We use the well-known AURORA-2 task (Hirsch and Pearce, 2000) as the platform for developing our modulation spectrum approach to noise-robust ASR. We will use the 'standard' back end for this task, i.e. a Viterbi decoder that finds the best path in a lattice spanned by the 179 states that result from representing 11 digit words by 16 states each, plus 3 states for representing non-speech. We expect that the effect of the additive noise is limited to a subset of the 135 output channels of the modulation spectrum analyser.

The major goal of this chapter is to introduce a novel approach to noise-robust ASR. The approach that we propose is novel in two respects: we use the 'raw' output of modulation frequency filters and we use Sparse Classification to derive state posterior probabilities from samples of the output of the modulation spectrum filters. We deliberately use unadorned implementations of both the modulation

spectrum analyser and the sparse coder, because we see a need for identifying what are the most important issues that are involved with a fundamentally different approach to representing speech signals and with converting the representations to state posterior estimates. In doing so we are fully aware of the risk that - for the moment - we will end up with word error rates (WERs) that are well above what is considered state-of-the-art (Bourlard et al., 1996b). Understanding the issues that affect the performance of our system most will allow us to propose a road map towards our final goal that combines advanced insight in what it is that makes human speech recognition so very robust against noise with improved procedures for automatic noise-robust speech recognition.

Our approach combines two novelties, *viz.* the features and the state posterior probability estimation. To make it possible to disentangle the contributions and implications of the two novelties, we will also conduct experiments in which we use conventional multi-layered perceptrons (MLPs) to derive state posterior probability estimates from the outputs of the modulation spectrum analyser. In section 3.4, we will compare the sparse classification approach with the results obtained with the MLP for estimating state posterior probabilities. This will allow us to assess the advantages of the modulation spectrum analyser, as well as the contribution of the sparse classification approach.

3.2 Method

3.2.1 Sparse classification frontend

The approach to noise-robust ASR that we propose in this chapter was inspired by Gemmeke et al. (2011b) and Gemmeke (2010), which introduced *Sparse Classification* as a technique for estimating the posterior probabilities of the lowest-level states in an ASR system. The starting point of their approach was a representation of noisy speech signals as overlapping sequences of up to 30 speech frames that together cover up to 300 ms intervals of the signals. Individual frames were represented as Mel-frequency energy spectra, because that representation conforms to the additivity requirement imposed by the sparse classification approach. SC is an exemplar-based approach. Handling clean speech requires the construction of an exemplar dictionary that contains stretches of speech signals of the same length

as the (overlapping) stretches that must be coded. The exemplars must be chosen such that they represent arbitrary utterances. For noisy speech a second exemplar dictionary must be created, which contains equally long exemplars of the additive noises. Speech is coded by finding a small number of speech and noise exemplars which, added together with positive weights, accurately approximate an interval of the original signal. The algorithms that find the best exemplars and their weights are called *solvers*; all solvers allow imposing a maximum on the number of exemplars that are returned with a weight > 0 so that it is guaranteed that the result is sparse. Different families of solvers are available, but some require that all coefficients in the representations of the signals and the exemplars are non-negative numbers. Least angle regression (Efron et al., 2004), implemented by means of a version of the *Lasso* solver, can operate with representations that contain positive and negative numbers.

The SC approach sketched above is interesting for two reasons. Sequences of short-time spectra implicitly represent a substantial part of the information in the modulation spectrum. That is certainly true if the sequences cover up to 300-ms signal intervals. In addition, in Hurmalainen et al. (2011b) it was shown that it is possible to convert the weights assigned to the exemplars in a SC system to the estimates of state probabilities, provided that the frames in the exemplars are assigned to states. The latter can be accomplished by means of a forced alignment of the database from which the exemplars are selected with the states that correspond to a phonetic transcription. In actual practice, the state labels are obtained by means of a forced alignment using a conventional hidden Markov model (HMM) recognizer.

The success of the SC approach in Gemmeke et al. (2011b) and Gemmeke (2010) for noise-robust speech recognition is attributed to the fact that the speech exemplars are characterized by peaks in the spectral energy that exhibit substantial continuity over time; the human articulatory system can only produce signals that contain few clear discontinuities (such as the release of stop consonants), while many noise types lack such continuity. Therefore, it is reasonable to expect that the modulation spectra of speech and noise are rather different, even if the short-time spectra may be very similar.

In this chapter we use the modulation spectrum directly to exploit the continuity constraints imposed by the speech production system. Since the modulation

spectrum captures information about the continuity of the speech signal in the low-frequency bands, there is no need for a representation that stacks a large number of subsequent time frames. Therefore, our exemplar dictionary can be created by selecting individual frames of the modulation spectrum in a database of labelled speech. As in Gemmeke et al. (2011b) and Gemmeke (2010), we will convert the weights assigned to the exemplars when coding unknown speech signals into estimates of the probability that a frame in the unknown signal corresponds to one of the states.

In Gemmeke et al. (2011b) and Gemmeke (2010) the conversion of exemplar weights into state probabilities involved an averaging procedure. A frame in an unknown speech signal was included in as many solutions of the solver as there were frames in an exemplar. In each position of a sliding window, an unknown frame is associated with the states in the exemplars chosen in that position. While individual window positions return a small number of exemplars and therefore a small number of possible states, the eventual set of state probabilities assigned to a frame is not very sparse. With the single-frame exemplars in the approach presented here, no such averaging is necessary or possible. The potential downside of relying on a single set of exemplars to estimate state probabilities is that it may yield overly sparse state probability vectors.

3.2.2 Data

In order to provide a proof of concept that our approach is viable, we used a part of the AURORA-2 database (Hirsch and Pearce, 2000). This database consists of speech recordings taken from the TIDIGITS corpus for which participants read sequences of digits (only using the words 'zero' to 'nine' and 'oh') with one up to seven digits per utterance. These recordings were then artificially noisified by adding different types of noise to the clean recordings at different signal-to-noise ratios. In this chapter we focus on the results obtained for test set A, i.e. the test set that is corrupted using the same noise types that occur in the multi-condition training set. We re-used a previously made state-level segmentation of the signals obtained by means of a forced alignment with a conventional HMM-based ASR system. These labels were also used to estimate the prior probabilities of the 179 states.

3.2.3 Feature extraction

The feature extraction process that we employ is illustrated in Figure 3.1. First, the (noisy) speech signal (sampling frequency F_s = 8 kHz) is analysed by a *gammatone filterbank* consisting of 15 band-pass filters with centre frequencies (F_c) spaced at one-third octave. More specifically, F_c = 125, 160, 200, 250, 315, 400, 500, 630, 800, 1,000, 1,250, 1,600, 2,000, 2,500, and 3,150 Hz, respectively. The amplitude response of an nth-order gammatone filter with centre frequency F_c is defined by

$$g(t) = a \cdot t^{n-1} \cdot \cos(2\pi F_c t + \phi) \cdot e^{-2\pi bt}. \tag{3.1}$$

With $b = 1.0183 \times (24.7 + F_c/9.265)$ and $n = 4$, this yields band-pass filters with equivalent rectangular bandwidth equal to 1 (Glasberg and Moore, 1990). Subsequently, the time envelope $e_i(t)$ of the ith filter output, x_i, is computed as the magnitude of the analytic signal

$$e_i(t) = \sqrt{x_i^2 + \hat{x}_i^2}, \tag{3.2}$$

with $\hat{x}_i$ the Hilbert transform of x_i. We assume that the time envelopes of the outputs of the gammatone filters are a sufficiently complete representation of the input speech signal. The frequency response of the gammatone filterbank is shown in the upper part at the left-hand side of Figure 3.1.

The Hilbert envelopes were low-pass filtered with a fifth-order Butterworth filter (*cf.* (3.3)) with cut-off frequency at 150 Hz and down-sampled to 400 Hz. The down-sampled time envelopes from the 15 gammatone filters are fed into another filterbank consisting of nine modulation filters. This so-called modulation filterbank is similar to the EPSM-filterbank as presented by Ewert and Dau (2000). In our implementation of the modulation filterbank, we used one-third-order Butterworth low-pass filter with a cut-off frequency of 1 Hz, and eight band-pass filters with centre frequencies of 2, 3, 4, 5, 6, 8, 10, and 16 Hz.[1]

[1]The software used for implementing the modulation frequency analyser was adapted from Matlab code that was kindly provided by Søren Jørgensen (Jørgensen and Dau, 2014). Some choices that are somewhat unusual in speech technology, such as the 400-Hz frame rate, were kept.

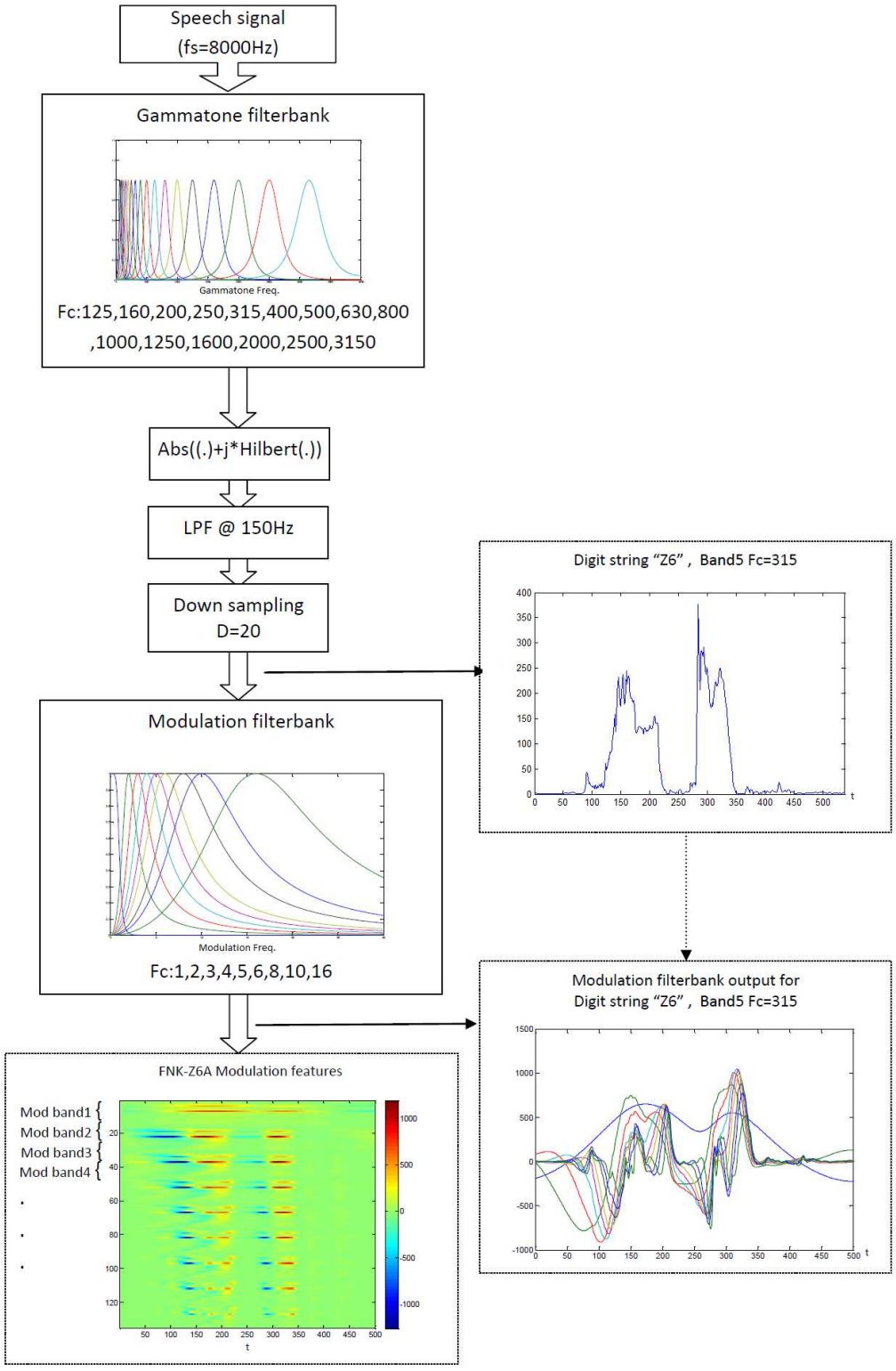


FIGURE 3.1: EMS feature extraction. The magnitude envelope of each of the 15 gammatone filters is decomposed into nine different modulation frequency bands. Thus, the speech is represented by $9 \times 15 = 135$-D feature vectors which are computed every 2.5 ms.

The frequency response of an nth-order low-pass filter with gain a and cut-off frequency F_c is specified by Rabiner and Gold (1975)

$$H(f) = \frac{a}{1.0 + (\frac{f}{F_c})^{2n}} \tag{3.3}$$

The complex-valued frequency response of a band-pass modulation filter with gain a, centre frequency F_c and quality factor $Q = 1$ is specified by

$$H(f) = \frac{a}{1.0 + jQ.\left(\frac{f}{F_c} - \frac{F_c}{f}\right)} \tag{3.4}$$

As an example, the upper panel at the right-hand side in Figure 3.1 shows the time envelope of the output of the gammatone filter with centre frequency at 315 Hz for the digit sequence 'zero six'. The frequency response of the complete filterbank, i.e. the sum of the responses of the nine individual filters, is shown in Figure 3.2. Due to the spacing of the centre frequency of the filters and the overlap of their transfer functions, we effectively give more weight to the modulation frequencies that are dominant in speech (Kanedera et al., 1999).

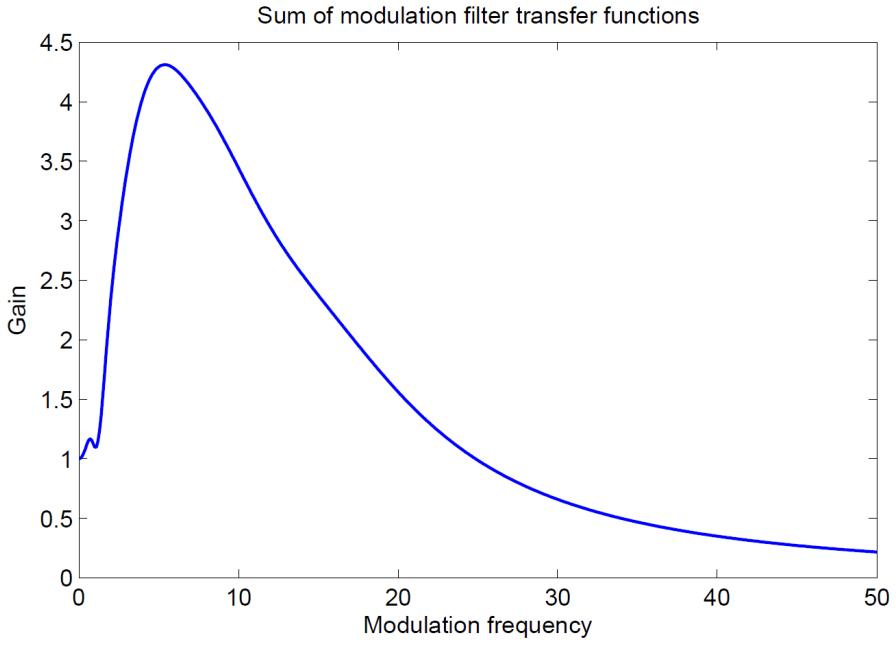


FIGURE 3.2: Sum of modulation transfer functions. The sum of the transfer functions of all modulation frequency filters gives a stronger weight to the frequencies that are known to be important for speech recognition (Kanedera et al., 1999).

The modulation frequency filterbank is implemented as a set of frequency domain filters. To obtain a frequency resolution of 0.1 Hz with the Hilbert envelopes sampled at 400 Hz, the calculations were based on Fourier transforms consisting of $4,001$ frequency samples. For that purpose we computed the complex-valued frequency response of the filters at 4,001 frequency points. An example of the ensemble of waveforms that results from the combination of the gammatone and modulation filterbank analysis for the digit sequence 'zero six' is shown in the lower panel on the right-hand side of Figure 3.1. The amplitudes of the $9 \times 15 = 135$ signals as a function of time are shown in the bottom panel at the left-hand side of Figure 3.1. The top band represents the lowest modulation frequencies (0 to 1 Hz) and the bottom band the highest (modulation filter with centre frequency $F_c = 16$ Hz).

We experimented with two different implementations of the modulation frequency filterbank, one in which we kept the phase response of the filters and the other in which we ignored the phase response and only retained the magnitude of the transfer functions. The results are illustrated in Figure 3.3 for clean speech and for the 5 dB signal-to-noise ratio (SNR) condition. From the second and third rows in that figure, it can be inferred that the linear phase implementation renders sudden changes in the Hilbert envelope as synchronized events in all modulation bands, while the full-phase implementation appears to smear these changes over wider time intervals. The (visual) effect is especially apparent in the right column, where the noisy speech is depicted. However, preliminary experiments indicated that the information captured in the 'visually noisy' full-phase representation could be harnessed by the recognition system: the full-phase implementation yields a performance increase in the order of 20% at the lower SNR levels compared with the performance of the linear phase implementation. However, the linear phase implementation works slightly better in clean and high SNR conditions (yielding ≈1% higher accuracies). This confirms the results of previous experiments in Moritz et al. (2011). Therefore, all results in this chapter are based on the full-phase implementation.

Another unsurprising observation that can be made from Figure 3.3 is that the non-negative Hilbert envelopes are turned into signals that have both positive and negative amplitude values. This will limit the options in choosing a solver in the SC approach to computing state posterior probabilities.

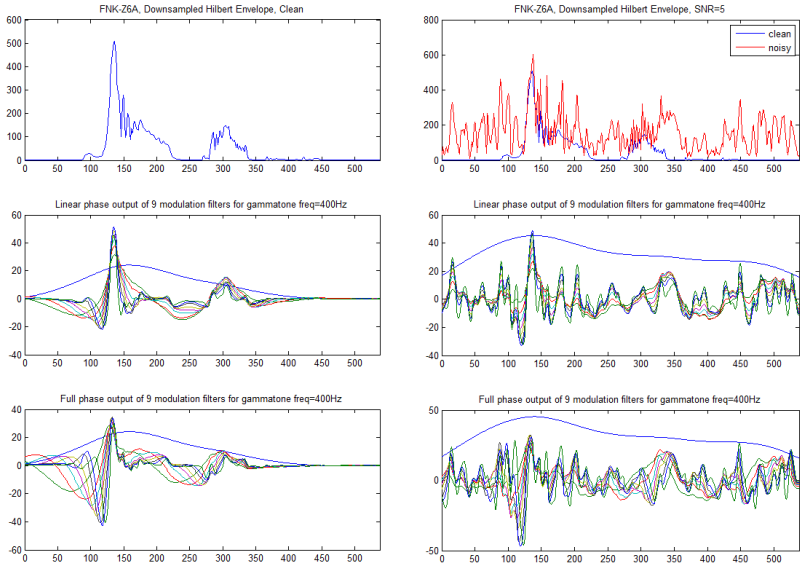


FIGURE 3.3: Full and linear phase features comparison. Top: spectrum envelope of 6th gammatone band for a sample utterance for clean (left) and 5-dB SNR (right). Middle: output of linear phase modulation filterbank. Bottom: output of full-phase modulation filterbank.

Figure 3.4 provides an extended view of the result of a modulation spectrum analysis of the utterance 'zero six'. The nine heat map representations in the lower left-hand part of Figure 3.1 are re-drawn in such a way that it is possible to see the similarities and differences between the modulation bands. The top panel in Figure 3.4 shows the output amplitude of the low-pass filter of the modulation filterbank. Subsequent panels show the amplitude of the outputs of the higher modulation band filters. It can be seen that overall, the amplitude decreases with increasing band number.

Speech and background noise tend to cover the same frequency regions in the short-time spectrum. Therefore, speech and noise will be mixed in the outputs of the 15 gammatone filters. The modulation filterbank decomposes each of the 15 time envelopes into a set of nine time-domain signals that correspond to different modulation frequencies. Generally speaking, the outputs of the lowest modulation frequencies are more associated with events demarcating syllable nuclei, while the higher modulation frequencies represent shorter-term events. We want to take advantage of the fact that it is unlikely that speech and noise sound sources with frequency components in the same gammatone filter also happen to overlap completely in the modulation frequency domain. Stationary noise would not affect

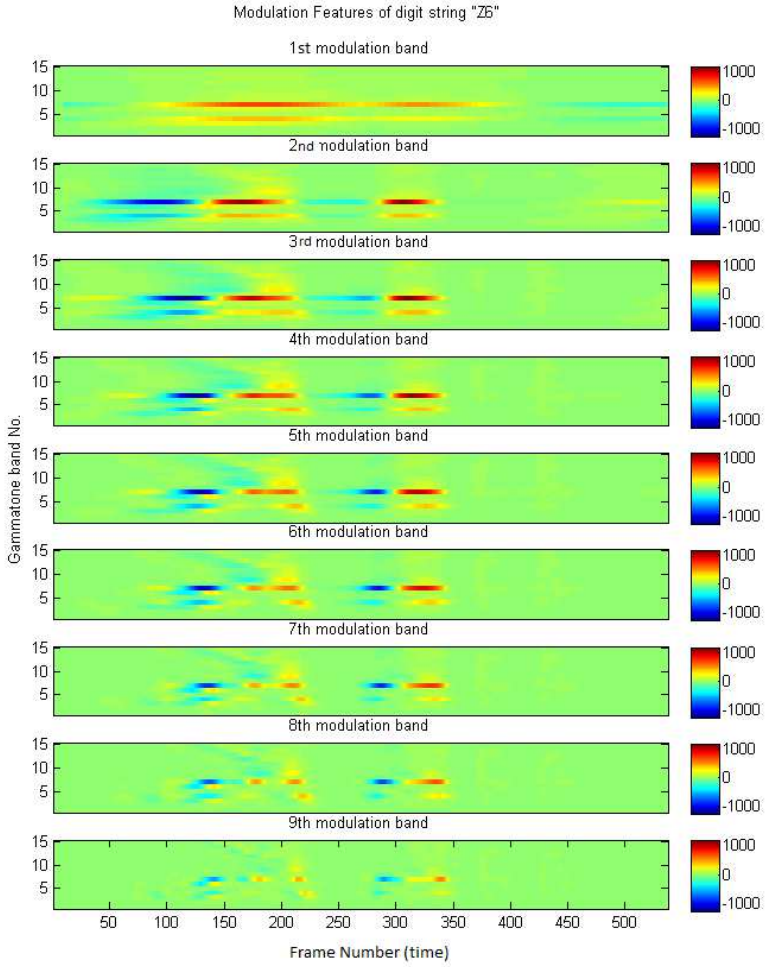


FIGURE 3.4: An extended view of the result of a modulation spectrum analysis of the utterance 'zero six'.

the output of the higher modulation frequency filters, while pulsatile noise should not affect the lowest modulation frequency filters. Therefore, we expect that many of the naturally occurring noise sources will show temporal variations at different rates than speech.

Although the EMS features capture short- and medium-time spectral dynamics, the information is encoded in a manner that might not be optimal for automatic pattern recognition purposes. Therefore, we decided to also create a feature set that encodes the temporal dynamics more explicitly. To that end we concatenated 29 frames (at a rate of one frame per 2.5 ms), corresponding to 29×2.5 = 72.5 ms; to keep the number of features within reasonable limits, we performed dimensionality reduction by means of linear discriminant analysis (LDA), with the 179 state labels as categories. The reference category was the state label of the middle frame

of a 29-frame sequence. The LDA transformation matrix was learned using the exemplar dictionary (*cf.*, section 3.2.4). The dimension of the feature vectors was reduced to 135, the same number as with single-frame features. To be able to investigate the effect of the LDA transform, we also applied an LDA transform to the original single-frame features. Here, the dimension of the transformed feature vector was limited to 135 (nine modulation bands in 15 gammatone filters).

3.2.4 Composition of exemplar dictionary

To construct the speech exemplar dictionary, we first encoded the clean train set of AURORA-2 with the modulation spectrum analysis system, using a frame rate of 400 Hz. Then, we quasi-randomly selected two frames from each utterance. To make sure that we had a reasonably uniform coverage of all states and both genders, 2×179 counters were used (one for each state of each gender). The counters were initialized at 48. For each selected exemplar, the corresponding counter was decremented by 1. Exemplars of a gender-state combination were no longer added to the dictionary if the counter became zero. A simple implementation of this search strategy yielded a set of $17,148$ exemplars, in which some states missed one or two exemplars. It appeared that 36 exemplars had a Pearson correlation coefficient of > 0.999 with at least one other exemplar. Therefore, the effective size of the dictionary is $17,091$.

We also encoded the four noises in the multi-condition training set of AURORA-2 with the modulation spectrum analysis system. From the output, we randomly selected $13,300$ frames as noise exemplars, with an equal number of exemplars for the four noise types.

When using LDA-transformed concatenated features, a new equally large set of exemplars was created by selecting sequences of 29 consecutive frames, using the same procedures as for selecting single-frame exemplars. In a similar vein, 29-frame noise exemplars were selected that were reduced to 135-D features using the same transformation matrix as for the speech exemplars.

3.2.5 The sparse classification algorithm

The use of sparse classification requires that it must be possible to approximate an unknown observation with a (positive) weighted sum of a number of exemplars. Since all operations in the modulation spectrum analysis system are linear and since the noisy signals were constructed by simply adding clean speech and noise, we are confident that the EMS representation does not violate additivity to such an extent that SC is rendered impossible. The same argument holds for the LDA-transformed features. Since linear transformations do not violate additivity, we assume that the transformed exemplars can be used in the same way as the original ones.

As can be seen in Figures 3.1 and 3.3, the output of the modulation filters contains both positive and negative numbers. Therefore, we need to use the Lasso procedure for solving the sparse coding problem, which can operate with positive and negative numbers (Efron et al., 2004). We are not aware of other solvers that offer the same freedom. Lasso uses the Euclidean distance as the divergence measure to evaluate the similarity of vectors. This raises the question whether the Euclidean distance is a suitable measure for comparing EMS feature vectors. We verified this by computing the distributions of the Euclidean distance between neighbouring frames and frames taken at random time distances of > 20 frames in a set of 100 randomly selected utterances. As can be seen from Figure 3.5, the distributions of the distances between neighbouring and distant frames hardly overlap. Therefore, we believe that it is safe to assume that the Euclidean distance measure is adequate.

Using the Euclidean distance in a straightforward manner implies that vector elements that have a large variance or large absolute values will dominate the result. Preliminary experiments showed that the modulation spectra suffer from this effect. It appeared that the difference between /u/ in *two* and /i/ in *three*, which is mainly represented by different energy levels in the 2,000-Hz region, was often very small because of the absolute values of the output of the modulation filters in the gammatone filters with centre frequencies of 2,000 and 2,500 Hz which were very much smaller than the values in the gammatone filters with centre frequencies up to 400 Hz. This effect can be remedied by using a proper normalization of the vector elements. After some experiments, we decided to equalize the variance in the gammatone bands. For this purpose we first computed the variance in all 135

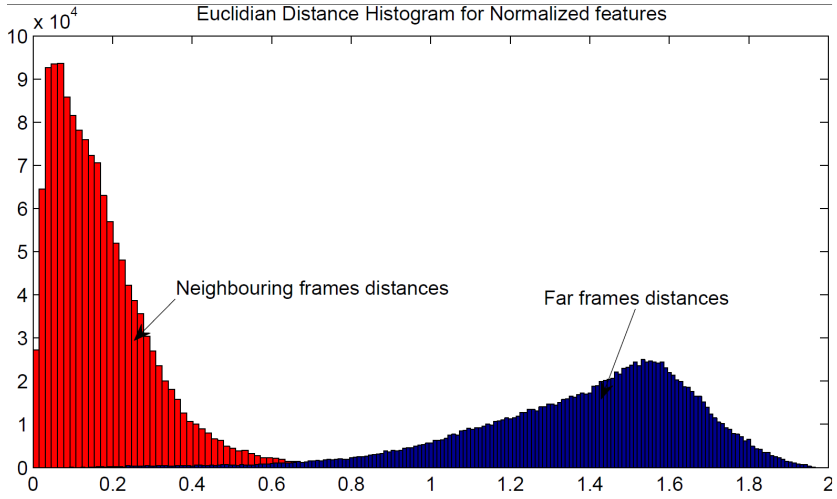


FIGURE 3.5: Distributions of the Euclidean distance between neighbouring (red) and distant (blue) **135**-D feature vectors.

modulation bands in the set of speech exemplars. Then, we averaged the variance over the nine modulation bands in each gammatone filter. The resulting averages were used to normalize the outputs of the modulation filters. The effect of this procedure on the representation of the output of the modulation filters is shown in Figure 3.6. This procedure reduced the number of /u/ - /i/ confusions by almost a factor of 3.

3.2.5.1 Obtaining state posterior estimates

The weights assigned to the exemplars by the Lasso solver must be converted to estimates of the probability that a frame corresponds to one of the 179 states. In the sparse classification system of Gemmeke et al. (2011b), weights of up to 30 window positions were averaged. In our SC system, we do not have a sliding window with heavy overlap between subsequent positions. We decided to use the weights of the exemplars that approximate individual frames to derive the state posterior probability estimates. In doing so, we simply added the weights of all exemplars corresponding to a given state. The average number of non-zero elements in the activation vector varied between 15.1 for clean speech and 6.5 at -5 dB SNR. Therefore, we may face overly sparse and potentially somewhat noisy state probability estimates. This is illustrated in Figure 3.7a for the digit sequence '3 6 7' in the 5 dB SNR condition. The traces of state probability estimates are not continuous (do not traverse all 16 states of a word) and they include activations of

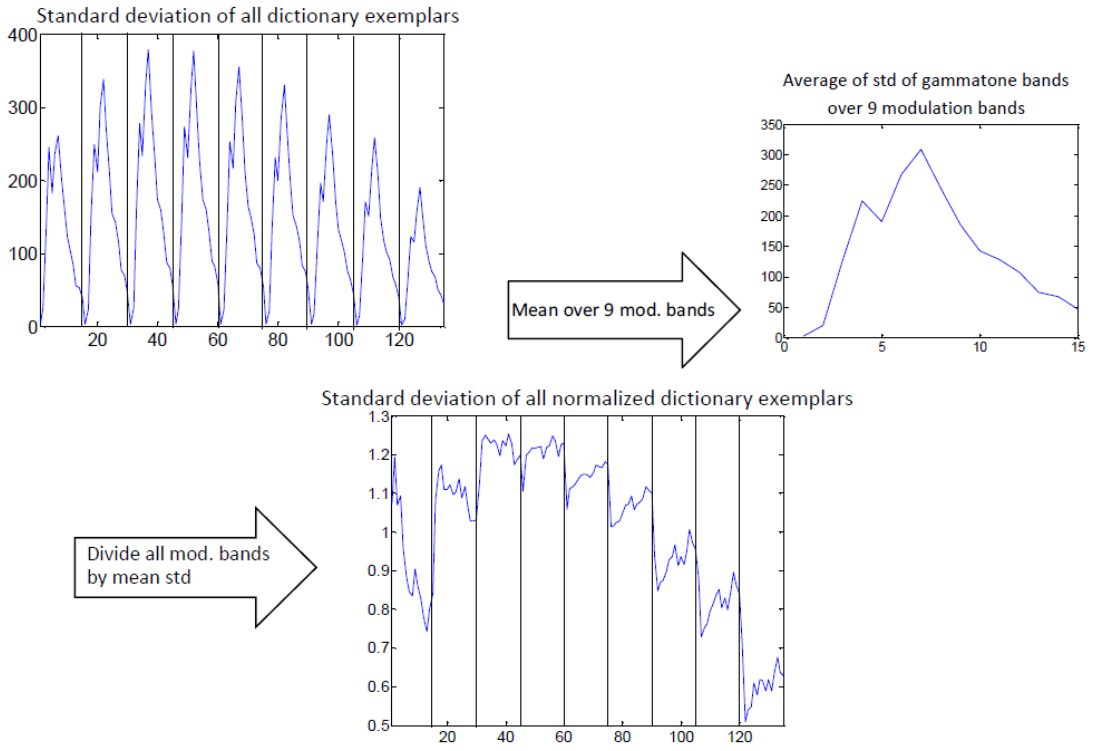
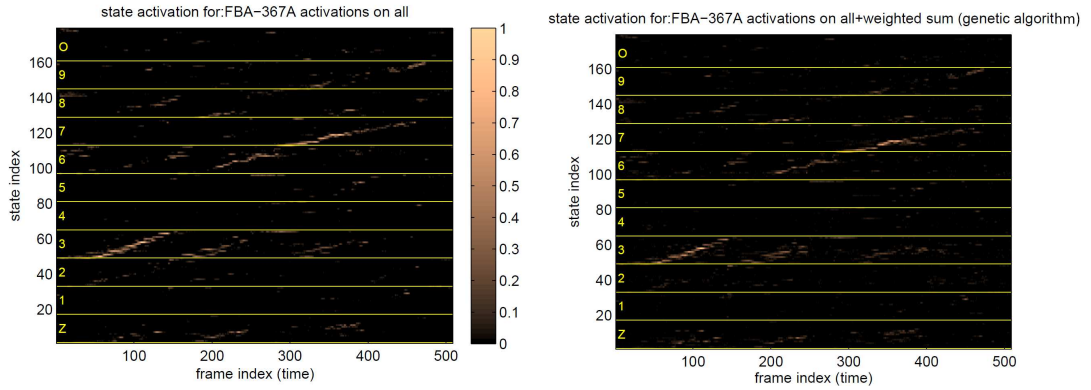


FIGURE 3.6: Normalization of the modulation filter outputs. Upper left: standard deviation of all 135 elements in the speech exemplars. Upper right: standard deviation in the gammatone filters averaged over all nine modulation filters. Lower panel: standard deviation of all 135 elements in the speech exemplars after normalization.

other states, some of which are acoustically similar to the states that correspond to the digit sequence.

3.2.6 Recognition based on combinations of individual modulation bands

Substantial previous research has investigated the possibility to combat additive noise by fusing the outputs of a number of parallel recognizers, each operating on a separate frequency band (*cf.*, Cerisara and Fohr (2001) for a comprehensive review). The general idea underlying this approach is that additive noise will only affect some frequency bands so that other bands should suffer less. The same idea has also been proposed for different modulation bands (Hermansky and Fousek, 2005a). In this chapter we also explore the possibility that additive noise does not affect all modulation bands to the same extent. Therefore, we will compare



(a) (b)

FIGURE 3.7: State probability traces for the digit sequence '3 6 7' at 5 dB SNR. **(a)** Traces obtained by using the activation weights of the full modulation spectrum exemplars only. The Viterbi decoder returns the incorrect sequence '3 3 7'. **(b)** Traces obtained from fusing the probability estimates obtained with the full modulation spectrum and the probability estimates obtained from nine modulation bands (weights obtained with a genetic algorithm). The Viterbi decoder now returns the correct sequence '3 6 7'.

recognition accuracies obtained when estimating state likelihoods using a single set of exemplars represented by 135-D feature vectors and the fusion of the state likelihoods estimated from the 135-D system and nine sets of exemplars (one for each modulation band) represented as 15-D feature vectors (for the 15 gammatone filters). The optimal weights for the nine sets of estimates will be obtained using a genetic algorithm with a small set of held-out training utterances. Also, combining state posterior probability estimates from ten decoders might help to make the resulting probability vectors less sparse.

3.2.7 State posteriors estimated by means of an MLP

In order to tease apart the contributions of the modulation frequency features and the sparse coding, we also conducted experiments in which we used a MLP for estimating the posterior probabilities of the 179 states in the AURORA-2 task. For this purpose we trained a number of networks by means of the QuickNet software package (Johnson et al., 2004). We trained networks on clean data only, as well as on the full set of utterances in the multi-condition training set. Analogously to Sun et al. (2012b), we used 90% of the training set, i.e. 7,596 utterances for training the MLP and the remaining 844 utterances for the cross-validation. To enable a fair comparison, we trained two networks, both operating on single frames.

The first network used frames consisting of 135 features; the second network used 'static' modulation frequency features extended with delta and delta-delta features estimated over a time interval of 90 ms, making for 405 input features. The delta and delta-delta features were obtained by fitting a linear regression on the sequence of feature values that span the 90-ms intervals. Actually, the 90-ms interval corresponds to the time interval covered by the perceptual linear prediction (PLP) features used in Sun et al. (2012b). There too, the static PLP features were extended by delta and delta-delta features, making for $9 \times 39 = 351$ input nodes.

3.3 Results

The recognition accuracies obtained with the 135-D EMS features are presented in the top part of Tables 3.1 and 3.2 for the SC-based system. The second and third rows of Table 3.2 show the results for the MLP-based system. Both tables also contain results obtained previously with conventional Mel-spectrum or PLP features. Note that the results in Table 3.1 pertain to a single-noise condition of test set A (subway noise), while Table 3.2 shows the accuracies averaged over all four noise types in test set A. In experimenting with the AURORA-2 task, it is a pervasive finding that the results depend strongly on the word insertion penalty (WIP) that is used in the Viterbi backend. A WIP that yields the lowest WER in the clean condition invariably gives a very high WER in the noisiest conditions. In this study we set aside a small development set, on which we searched the WIP that gave the best results in the conditions with SNR $\leq$ 5 dB; in these conditions the best performance was obtained with the same WIP value. Inevitably, this means that we will end up with relatively bad results in the cleanest conditions. Unfortunately, there is no generally accepted strategy for selecting the 'optimal' WIP. Since different authors make different (and not always explicit) decisions, detailed comparisons with results reported in the literature are difficult. For this study this is less of an issue, since we are not aiming at outperforming previously published results.

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB
Sys1 (single frame)	90.51	91.00	89.53	87.69	83.76	76.76	65.31
Sys2 (single frame) (LDA transformed)	89.19	89.62	87.57	83.54	76.51	62.57	36.91
Sys3 (29 frames) (LDA transformed)	87.50	88.70	87.41	85.42	77.62	59.41	27.85
Sys4 (9 bands-GA)	89.71	90.57	89.28	87.41	84.13	77.71	63.83
Sparse coding (Gemmeke, 2010) 5-frame exemplars	93.12	90.18	87.22	82.62	72.64	56.31	34.57
Sparse coding (Gemmeke, 2010) 30-frame exemplars	93.21	91.86	91.53	89.62	87.47	80.01	61.61

TABLE 3.1: Accuracy for five systems on noise type 1 (subway noise) of test set A. Sys1, 135-D vectors; Sys2, LDA-transformed 135-D vectors of Sys1; Sys3, LDA-transformed 29×135-D vectors of 29 consecutive frames; Sys4, Sys1 plus nine recognizers operating on 15-D vectors, weights obtained from a genetic algorithm. Recognition results for noise type 1 using the sparse coding approach (Gemmeke et al., 2011b; Gemmeke, 2010) using 5 and 30 frame windows are included for comparison in the bottom part.

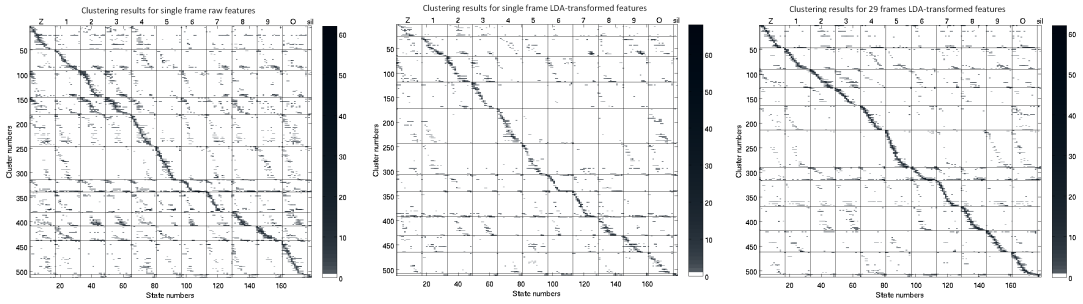
3.3.1 Analysing the features

To better understand the EMS features, we carried out a clustering analysis on the exemplars in the dictionary, using k-means clustering. We created 512 clusters using the scikit-learn software package (Pedregosa et al., 2011). We then analysed the way in which clusters correspond to states. The results of the analysis of the raw features are shown in Figure 3.8a. The horizontal axis in the figure corresponds to the 179 states, and the vertical axis to cluster numbers. The figure shows the association between clusters and states. It can be seen that the exemplar clusters do associate to states, but there is a substantial amount of 'confusions'. Figure 3.8b shows the result of the same clustering of the exemplars after applying

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB
EMS features sparse coding 1-frame exemplar (Sys1)	90.62	90.87	89.90	88.17	84.46	76.83	59.65
EMS features MLP 135 input nodes multi-condition	96.93	96.66	95.84	94.07	87.14	68.05	35.46
EMS features + Δ + ΔΔ MLP 405 input nodes multi-condition	97.71	97.36	96.74	95.08	89.79	70.58	34.55
PLP + Δ and ΔΔ MLP 351 input nodes (Sun et al., 2012b) multi-condition	99.08	98.89	98.45	96.89	91.80	72.80	35.67
Mel features sparse coding (Gemmeke, 2010) 5-frame exemplars	93.43	90.94	89.06	84.57	75.91	58.20	32.57
Mel features sparse coding (Gemmeke, 2010) 30-frame exemplars	93.68	92.53	92.02	90.78	88.01	78.93	57.11

TABLE 3.2: Accuracies averaged over all noise types in test set A obtained with Sys1 (SC system operating on 135-D EMS features), MLP classifiers (on same features without and with Δs and ΔΔs), MLP classifier on PLP features with Δs and ΔΔs (Sun et al., 2012b), SC classifier on Mel spectra (Gemmeke, 2010) using 5- and 30-frame windows, respectively.

an LDA transform to the exemplars, keeping all 135 dimensions. It can be seen that the LDA-transformed exemplars result in clusters that are substantially purer. Figure 3.8c shows the results of the same clustering on the 135-D features obtained from the LDA transform of sequences of 29 subsequent frames. Now, the cluster purity has increased further.



(a) (b) (c)

FIGURE 3.8: Clustering results. **(a)** Single-frame raw features. **(b)** Single-frame LDA-transformed features. **(c)** The 29-frame LDA-transformed features.

Although cluster purity does not guarantee high recognition performance, from Tables 3.1 and 3.2 it can be seen that the EMS features appear to capture substantial information that can be exploited by two very different classifiers.

3.3.2 Results obtained with the SC system

Table 3.1 summarizes the recognition accuracies obtained with six different systems, all of which used the SC approach to estimate state posterior probabilities. Four of these systems use the newly proposed EMS features, while the remaining two describe the results using Mel-spectrum features as obtained in research done by Gemmeke (Gemmeke, 2010).

From the first three rows of Table 3.1, it can be seen that estimating state posterior probabilities from a single frame of a modulation spectrum analysis by converting the exemplar weights obtained with the sparse classification system already yields quite promising results. Indeed, from a comparison with the results obtained with the original SC system using five-frame stacks in Gemmeke (2010), it appears that the EMS features outperform stacks of five Mel-spectrum features in all but one condition. The conspicuous exception is the clean condition, where the performance of *Sys1* is somewhat disappointing. Our *Sys1* performs worse

than the system in Gemmeke (2010) that used 30-frame exemplars. From the first and second rows, it can be inferred that transforming the features such that the discrimination between the 179 states is optimized is harmful for all conditions. Apparently, the transform learned on the basis of 17.148 exemplars does not generalize sufficiently to the bulk of the feature frames. In section 3.4 we will propose an alternative perspective that puts part of the blame on the interaction between LDA and SC.

3.3.2.1 The representation of the temporal dynamics

In Gemmeke et al. (2011b) and Gemmeke (2010) the recognition performance in AURORA-2 was compared for exemplar lengths of $5, 10, 20$, and 30 frames. For clean speech, the optimal exemplar length was around ten frames and the performance dropped for longer exemplars; at SNR = -5 dB, increasing exemplar length kept improving the recognition performance and the optimal length found was the longest that was tried (i.e. 30). Longer windows correspond with capturing the effects of lower modulation frequencies. The trade-off between clean and very noisy signals suggests that emphasizing long-term continuity helps in reducing the effect of noises that are not characterized by continuity, but using 300-ms exemplars may not be optimal for covering shorter-term variation in the digits. From the two bottom rows in Table 3.1, it can be seen that going from 5-frame stacks to 30-frame stacks improved the performance for the noisiest conditions very substantially. From the second and third rows in that table, it appears that the performance gain in our system that used 29-frame features (covering 72.5 ms) is nowhere near as large. However, due to the problems with the generalizability of the LDA transform that we already encountered in *Sys2*, it is not yet possible to draw conclusions from this finding.

A potentially important side effect of using exemplars consisting of 30 subsequent frames in Gemmeke et al. (2011b) and Gemmeke (2010) was that the conversion of state activations to state posterior probabilities involved averaging over 30 frame positions. This diminishes the risk that a 'true' state is not activated at all. Our system approximates a feature frame as just one sum of exemplars. If an exemplar of an 'incorrect' state happens to match best with the feature frame, the Lasso procedure may fill the gap between that exemplar and the feature frame with completely unrelated exemplars. This can cause gaps in the traces in the state

probability lattice that represent the digits. This effect is illustrated in Figure 3.7a, which shows the state activations over time of the digit sequence '3 6 7' at 5 dB SNR for the state probabilities in *Sys1*. The initial fricative consonants /θ/ and /s/ and the vowels /i/ and /ɪ/ in the digits '3' and '6' are acoustically very similar. For the second digit in the utterance, this results in somewhat grainy, discontinuous, and largely parallel traces in the probability lattice for the digits '3' and '6'. Both traces more or less traverse the sequence of all 16 required states. The best path according to the Viterbi decoder corresponds to the sequence '3 3 7', which is obviously incorrect.

3.3.2.2 Results based on fusing nine modulation bands

In *Sys1, Sys2*, and *Sys3*, we capitalize on the assumption that the sparse classification procedure can harness the differences between speech and noise in the modulation spectra without being given any specific information. In Cerisara and Fohr (2001) it was shown that it is beneficial to 'help' a speech recognition system in handling additive noise by fusing the results of independent recognition operations on non-overlapping parts of the spectrum. The success of the multi-band approach is founded in the finding that additive noise does not affect all parts of the spectrum equally severely. Recognition on sub-bands can profit from superior results in sub-bands that are only marginally affected by the noise. Using EMS features, we aim to exploit the different temporal characteristics of speech and noise, which are expected to have different effects in different modulation bands. Therefore, we conducted an experiment to investigate whether combining the output of nine independent recognizers, each operating on a different modulation frequency band, will improve recognition accuracy. In each modulation frequency band, we have the output of all 15 gammatone filters; therefore, each modulation band 'hears' the full 4-kHz spectrum. The experiment was conducted using the part of test set ***A*** that is corrupted by subway noise.

In our experiments we opted for fusion at the state posterior probability level: We constructed a single-state probability lattice for each utterance by means of a weighted sum of the state posteriors obtained from the individual SC systems. In all cases we fused the probability estimates of *Sys1*, which operates with 135-D exemplars with nine sets of state posteriors from SC classifiers that each operate on 15-D exemplars. *Sys1* was always given a weight equal to 1. The weights for

the nine modulation band classifiers were obtained using a genetic algorithm that optimized the weights on a small development set. The weights and WIP that yielded the best results in the SNR conditions ≤5 dB were applied to all SNR conditions. The set of weights is shown in Table 3.3.

From row 4 (*Sys4*) in Table 3.1, it can be seen that fusing the state likelihood estimates from the nine individual modulation filters with the state likelihoods from the full modulation spectrum deteriorates the recognition accuracy for all but two SNRs. From Table 3.3 it appears that the Genetic Algorithm returns very small weights for all nine modulation bands. This strongly suggests that the individual modulation bands are not able to highlight specific information that is less easily seen in the complete modulation spectrum.

A potentially very important concomitant advantage of fusing the probability estimates from the 135-D system and the nine 15-D systems is that the fusion process may make the probability vectors less sparse, thereby reducing the risk that incorrect states are being promoted. This is illustrated in Figure 3.7b, where it can be seen that the state probability traces obtained from the fusion of the full 135-D system and the weighted sub-band systems suffer less from competing 'ghost traces' of acoustically similar competitors that traverse all 16 states of the incorrect digit: Due to the lack of consensus between the multiple classifiers, the trace for the incorrect digit '3', which is clearly visible in Figure 3.7a, has become less clear and more 'cloud-like' in Figure 3.7b. As a consequence, the digit string is now recognized correctly as '3 6 7'. However, from the results in Table 3.1, it is clear that on average the impact of making the probability vectors less sparse by means of fusing modulation frequency sub-bands is negligible.

F_c (Hz)	0	2	3	4	5	6	8	10	16
GA	–0.0172	–0.0921	0.0001	–0.0103	–0.223	–0.0336	–0.0072	–0.0625	0.201

TABLE 3.3: Weights obtained for combining the 15 gammatone filterbands in the multi-stream analysis. GA, weights obtained with a genetic algorithm.

3.3.3 Results obtained with MLPs

We trained four MLP systems for computing state posterior probabilities on the basis of the EMS features, two using only clean speech and two using the multi-condition training data. We increased the number of hidden nodes, starting with 200 hidden nodes up to $1,500$ nodes. In all cases the eventual recognition accuracy kept increasing, although the rate of increase dropped substantially. Additional experiments showed that further increasing the number of hidden nodes no longer yields improved recognition results. For each number of hidden nodes, we also searched for the WIP that would provide optimal results for the cross-validation set (*cf.* section 3.2.7). We found that the optimal accuracy in the different SNR conditions was obtained for quite different values of the WIP. Training on multi-condition data had a slight negative effect on the recognition accuracy in the clean condition, compared to training on clean data only. However, as could be expected, the MLPs trained on clean data did not generalize to noisy data.

Table 3.2 shows the results obtained with SC systems operating on modulation spectrum (EMS) and Mel-spectrum features and the MLP-based systems trained with multi-condition data. It can be seen that adding Δ and $\Delta\Delta$ features to the EMS features increases performance somewhat, but by no means to the extent that adding Δ and $\Delta\Delta$ features improves performance with Mel-spectrum or PLP features (Mlouka and Liénard, 1974; Furui, 1981).

The two systems that used EMS features perform much worse on clean speech than the MLP-based system that used nine adjacent 10-ms PLP + Δ + $\Delta\Delta$ features (Sun et al., 2012a). This suggests that the EMS features fail to capture part of the dynamic information that is represented by the speed and acceleration features derived from PLPs. Interestingly, that information is not restored by adding the regression coefficients obtained with stacks of modulation frequency features. In the noisier conditions, the networks trained with modulation frequency features derived from the multi-condition training data approximate the performance of the stacks of nine extended PLP features.

3.4 Discussion

In this chapter we introduced a basic implementation of a noise-robust ASR system that uses the modulation spectrum, instead of the short-time spectrum to represent noisy speech signals, and sparse classification to derive state probability estimates from time samples of the modulation spectrum. Our approach differs from previous attempts to deploy sparse classification for noise-robust ASR. The first difference is the use of the modulation spectrum and the second is that the exemplars in our system are constituted by individual frames, rather than by (long) sequences of adjacent frames in Gemmeke et al. (2011b) and Gemmeke (2010), which needed such sequences to effectively cover essential information about continuity over time that comes for free in the modulation spectrum, where individual frames capture information about the dynamic changes in the short-time spectrum. Our unadorned implementation yielded recognition accuracies that are slightly below the best results in Gemmeke et al. (2011b) and Gemmeke (2010), but especially the fact that our system yielded higher accuracies in the -5 dB SNR condition than their systems with exemplars with a length of 50 ms corroborates our belief that we are on a promising track towards a novel approach to noise-robust ASR. Although all results are based on a combination of feature extraction and posterior state probability estimation, we will discuss the features and the estimators separately - to the extent possible.

3.4.1 The features

In designing the modulation spectrum analysis system, a number of decisions had to be made about implementation details. Although we are confident that all our decisions were reasonable (and supported by data from the literature), we cannot claim that they were optimal. Most data in the literature on modulation spectra are based on perception experiments with human subjects, but more often than not these experiments use auditory stimuli that are very different from speech. While the results of those experiments surely provide guidance for ASR, it may well be that the automatic processing aimed at extracting the discriminative information is so different from what humans do that some of our decisions are sub-optimal. Our gammatone filterbank contains 15 one-third octave filters, which have a higher resolution in the frequencies $<$ 500 Hz than the Mel filterbank that is used in

most ASR systems. However, initial experiments in which we compared our one-third octave filterbank with a filterbank consisting of 23 Mel-spaced gammatone filters, spanning the frequency range of 64 to 3,340 Hz did not show a significant advantage of the latter over the former. From the speech technology's point of view, this may seem surprising because the narrow-band filters of the one-third octave filterbank in the low frequencies may cause interactions with fundamental frequency, while the relatively broad filters in the higher frequencies cannot resolve formants. But from an auditory system's point of view, there is no such surprise, since one-third octave filters are compatible with most, if not all, outcomes of psycho-acoustic experiments. This is also true for experiments that focused on speech intelligibility (Drullman et al., 1994).

For the modulation filterbank, it also holds that the design is partly based on the results of perception experiments (Jørgensen and Dau, 2011). Our modulation frequency analyser contained filters with centre frequencies ranging from 0 to 16 Hz. From Kanedera et al. (1999) it appears that the modulation frequency range of interest for ASR is limited to the 2- to 16-Hz region. Therefore, here too we must ask whether our design is optimal for ASR. It might be that the spacing of the modulation filters in the frequency band that is most important for human speech intelligibility is not optimal for automatic processing. However, as with the gammatone filters, it is not evident why a different spacing should be preferred. It might be necessary to treat modulation frequencies $\leq$1 Hz, which are more likely to correspond to the characteristics of the transmission channel, different than modulation frequencies that might be related to articulation. One might think that the very low modulation frequencies would best be discarded completely in the AURORA-2 task, where transmission channel characteristics do not play a role. However, experiments in which we did just that yielded substantially worse results. Arguably, the lowest modulation frequencies help in distinguishing time intervals that contain speech from time intervals that contain only silence or background noise. We decided to not include modulation filters with centre frequencies >16 Hz. This implies that we ignore almost all information related to the periodicity that characterizes many speech sounds. However, it is well known that the presence of periodicity is a powerful indicator of the presence of speech in noisy signals and also, in case the background noise consists of speech from one or more interfering speakers, a powerful means to separate the target speech from the background

speech. In future experiments we will investigate the possibility of adding explicit information about the harmonicity of the signals to the feature set.

The experiments with the MLP classifiers for obtaining state posterior probabilities from the EMS features confirm that the EMS features capture most of the information that is relevant for speech decoding. Still, the WERs obtained with the MLPs were always inferior to the results obtained with stacks of nine conventional PLP features that include Δ and $\Delta\Delta$ features, especially in the cleanest SNR conditions. Although the EMS features are performing quite well in noisy conditions, in cleaner conditions their performance is worse than the classical PLP features. Adding Δs and $\Delta\Delta$s, computed as linear regressions over 90 ms windows, to the EMS features does not improve performance nearly as much as adding speed and acceleration to MFCC or PLP features. This suggests that our EMS features are suboptimal with respect to describing the medium-term dynamics of the speech signal. The time windows associated with the modulation frequency filters with the lowest centre frequencies is larger than 500 ms. As a consequence, time derivatives computed over a window of 90 ms for these slowly varying filter outputs is not likely to carry much additional information. We suspect that the features in the lowest modulation bands play too heavy a role. If we want to optimally exploit the redundancy in the different modulation frequency channels when part of them gets obscured by noise, information about relevant speech events (such as word or syllable onsets and offsets) should ideally be represented equally well by their temporal dynamics in all channels.

Perhaps the most striking difference between the auditory model used in this study and the model proposed in Dau et al. (1996) is the absence of the adaptation/compression network between the gammatone filters and the modulation frequency filters. Preliminary experiments in which we applied tenth root compression to the output of the modulation filters (rather than the gammatone filters) already showed a substantial beneficial effect. The additional high-pass filtering that is performed in the compression/adaptation network (which should only be applied to the output of the gammatone filters) is expected to have a further beneficial effect in that it implements the medium-term dynamics that we seem to be missing at the moment. Including the adaptation stage is also expected to enhance the different dynamic characteristics of speech and many noise types in the modulation frequency bands. If this expectation holds, the absence of a proper adaptation network might explain the failure of the nine band fusion system.

3.4.2 The classifiers

Visual inspection of traces of state activations as a function of time obtained with the SC system suggested that the similarity between adjacent feature vectors was much higher than the similarity between adjacent state activation vectors. Figure 3.9 shows scatter plots of the relation between the similarity between adjacent feature vectors and the corresponding state probability vectors. It can be seen that the Pearson correlation coefficient between adjacent feature frames is very high, which is what one would expect, given the high sampling rate. It is also evident, and expected, that the variance increases as the SNR decreases. However, the behaviour of the state probability vectors is quite different. While for part of the adjacent vectors it holds that they are very similar (the pairs with a similarity close to one, represented by the points in the upper right-hand corner of the panels), it can be seen that there is a substantial proportion of adjacent state probability vectors that are almost orthogonal. We believe that this discrepancy is related to the difference between sparse coding (reconstruction of an observed modulation spectrum in terms of a linear combination of exemplars), what it is that the Lasso solver does, and sparse *classification* (estimating the probability of the HMM state underlying the observed modulation spectrum), which is our final goal. The frames that represent an unknown (noisy) speech signals are all decoded individually; for each frame the Lasso procedure starts from scratch. If occasionally a speech atom related to an incorrect state or an atom from the noise dictionary happens to match best with an input frame, this can have a very large impact on the resulting state activation vector. Lasso can turn a close similarity between an input frame and exemplars related to the true state at the feature level into a close-to-zero probability of the 'correct' state in the probability vector because an exemplar related to another state (or noise) happened to match slightly better.

The substantial deterioration of the recognition performance with LDA-transformed features came as a surprise, not in the last place because we have seen that cluster purity increases after LDA transform. The fact that we see a negative effect of the transform already for clean speech suggests that the transformation matrix learned from the exemplar dictionary does not generalize well to the continuous speech data. While the correlation between the raw features in adjacent frames was very close to one, the average Pearson correlation coefficient between adjacent

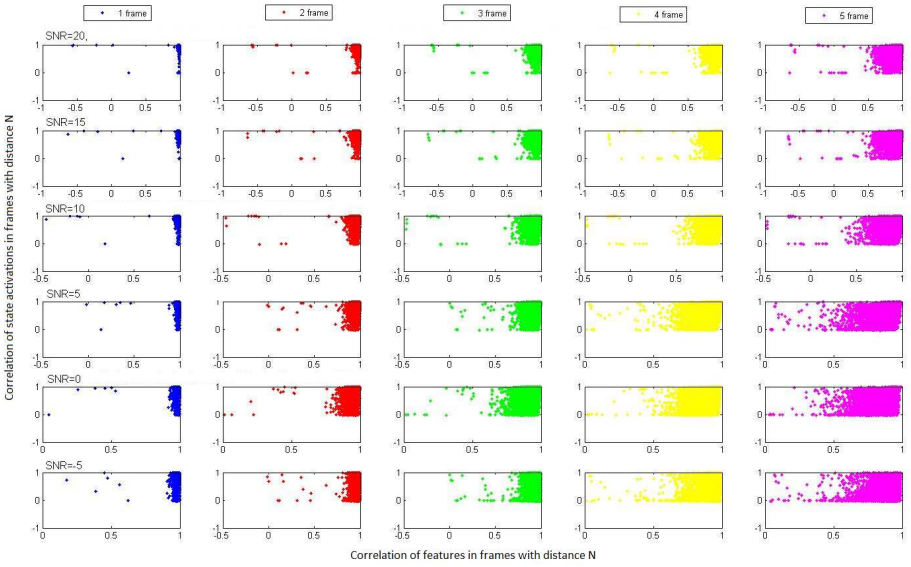


FIGURE 3.9: Relationship between the similarity between adjacent time frames and the corresponding adjacent state activation vectors. Correlations are computed using ten randomly selected utterances. Rows, SNR conditions. Columns, time distance between frames.

frames dropped to about 0.75 after the LDA transform. The LDA transformation maximizes the differences among the 179 states, regardless of whether states are actually very similar or not. Distinguishing adjacent states in the digit word 'oh' is equally important as distinguishing the eighth state of oh from the first state of 'seven'. Exaggerating the differences between adjacent frames, because these may relate to different states, is likely to aggravate the risk that Lasso returns high activations for an incorrect state because an exemplar assigned to that state happens to fit the frame under analysis best. In addition, the LDA transform affects the relations between the distributions of the features. Because we believe that the feature normalization applied to the raw EMS features yielded the best performance since it conforms with the mathematics in Lasso, we applied the same normalization to the LDA-transformed features. We did not (yet) check whether a different normalization could improve the results. The comparison between the single-frame LDA-transformed and 29-frame features that are reduced to 135-D features by means of an LDA-transform shows that adding a more explicit representation of the time context only improves the recognition accuracy in the 10 and 5 dB SNR conditions; in all other conditions, the results obtained with single-frame features are better. We believe that this finding is related to the difficulty of representing medium-term speech dynamics in the present form of the EMS

features.

We experimented with LDA in order to be able to explicitly include additional information about temporal dynamics. In the present implementation, with its 400-Hz frame rate, a stack of 29 adjacent frames covers a time interval of 72.5 ms, resulting in 3,915-D feature vectors. However, the 400-Hz frame rate does not seem to be necessary. Preliminary experiments with low-pass filtering the Hilbert envelopes of the outputs of the gammatone filters with a cut-off frequency of 50 Hz and a frame rate of 100 Hz yielded equal WER results. This opens the possibility of covering time spans of about 70 ms by concatenating only nine frames. However, an experiment in which we decoded the clean speech with exemplars consisting of nine subsequent 10-ms frames did not yield accuracies better than what we had obtained with single-frame features. This corroborates our belief that the medium-term dynamics is not sufficiently captured by our EMS features.

The success of the MLP classifiers that is apparent from Table 3.2 shows that sparse classification is not the only way for estimating state posterior probabilities from EMS features. In fact, the MLP classifier yielded consistently better results than the SC classifier in the SNR conditions covered by the training data. However, in the 0 and -5 dB SNR conditions, which are not present in the multi-condition training, the SC classifier yielded better performance. This raises the question whether it is possible to add supervised learning to the design of an SC-based system without sacrificing its superior generalization to unseen conditions.

In Gemmeke et al. (2011b) and Gemmeke (2010) it is mentioned that they failed to improve the performance of their sparse coding systems by machine learning techniques in the construction of the exemplar dictionaries. However, the cause of the failure was not explained. It may well be that the situation with single-frame EMS exemplars is different from 30-frame Mel-spectrum exemplars so that clever dictionary learning might be beneficial. We have started experiments with fast dictionary learning along the lines set out in Mairal et al. (2009a). Our first results suggest that there are two quite different issues that must be tackled. The first issue relates to the cost function used in creating the optimal exemplars. Conventional approaches to dictionary learning use the difference between unknown frames and their approximation as a weighted sum of exemplars as the criterion to minimize. While this criterion is obviously valid in sparse *coding* applications, it is not the criterion of choice in sparse *classification*. In the latter application,

the exemplars carry information about the states (the classes) that they represent, and this information should enter into the cost function, for example, in the form of the requirement that individual exemplars are promoted for frames that do correspond to a certain state (or set of acoustically similar states).

The second issue is that the mapping from state activations returned by some *solver* to state posterior probabilities is less straightforward than was implemented in Gemmeke et al. (2011b) and Gemmeke (2010) and in this study. There is a need for including some learning mechanisms that can find the optimal mapping from a complete set of state activations to a set of state posteriors. It is quite possible that there will be interactions between enhanced dictionary learning and learning the mapping from activations to probabilities. The challenge here is to find strategies that do not fall into the trap that we have seen in our experiments with MLPs, *viz.*, that the eventual performance increases in the conditions for which training material was available but at the cost of a diminished generalization to unseen conditions.

An issue that surely needs further investigation in the construction of the dictionary is the selection of the noise exemplars. So far, noise exemplars were extracted quasi-randomly from the four noise types that were used in creating the multi-condition training set in AURORA-2. It is quite likely that the collection of noise exemplars is much more compactly distributed in the feature space than the speech exemplars because the variation in the noise signals is less than in the speech signals. The generalization to other noise types can be improved by sampling the exemplars from a wider range of noises, for example all noise types that are available in the NOISEX CD-ROM (Varga and Steeneken, 1993). However, we think that the most important issue in the construction of the noise exemplar dictionary is the need for avoiding overlap between noise and speech exemplars. In the Lasso procedure, it is difficult - if not impossible - to enforce a preference for speech atoms over noise atoms. If a noise exemplar that is very similar to a speech exemplar happens to fit best, this may give rise to suppressing relevant speech information. It might be beneficial to not simply discard all noise exemplars activations but rather to include these, along with the activations of the speech exemplars, in a procedure that learns the mapping from activations to state posteriors that optimizes recognition performance. An approach in which all activations are used in estimating the eventual posterior probabilities would be especially important in cases where noise and speech are difficult to distinguish in

terms of spectro-temporal properties, such as in babble noise or if the 'noise' consists of competing speakers. These cases will surely require additional processing, for example, aimed at tracking continuity in pitch, in addition to continuity in the modulation spectrum.

3.5 Conclusions

In this chapter we presented a novel noise-robust ASR system that uses the modulation spectrum in combination with a sparse coding approach for estimating state probabilities. Importantly, in its present implementation, the system does not involve any form of learning/training. The best recognition accuracies obtained with the novel system are slightly below the results that have been obtained with conventional engineering systems. We have also sketched several research lines that hold the promise of improving the results and, at the same time, to advance our knowledge of those aspects of the human auditory system that are most important for ASR. We have shown that the output of a modulation spectrum analyser that does not involve any form of conversion to the equivalent of a short-time power spectrogram is able to exploit the spectro-temporal continuity constraints that are typical for speech and which are a prerequisite for noise robust ASR. However, we also found that the representation of medium-term dynamics in the output of the modulation spectrum analyser must be improved. With respect to the sparse coding approach to estimate state posterior probabilities, we have found that there is a fundamental distinction between sparse coding, where the task is to find the optimal representation of an unknown observation in a very large dimensional space, and sparse classification, where the task is to obtain the best possible estimates of the posterior probability that an unknown observation belongs to a specific class. In this context one challenge for future research is developing a procedure for dictionary learning that uses state posterior probabilities, in addition to or rather than reconstruction error, as the cost function. The second challenge is finding a procedure for learning a mapping from state activations to state posterior probabilities that provides the same excellent generalization to unseen conditions that has been found with sparse coding.

Chapter 4

Human-inspired Modulation Frequency Features for Noise-robust ASR

This study investigates a computational model that combines a frontend based on an auditory model with an exemplar-based sparse coding procedure for estimating the posterior probabilities of sub-word units when processing noisified speech. Envelope modulation spectrogram (EMS) features are extracted using an auditory model which decomposes the envelopes of the outputs of a bank of gammatone filters into one lowpass and multiple bandpass components. Through a systematic analysis of the configuration of the modulation filterbank, we investigate how and why different configurations affect the posterior probabilities of sub-word units by measuring the recognition accuracy on a semantics-free speech recognition task. Our main finding is that representing speech signal dynamics by means of multiple bandpass filters typically improves recognition accuracy. This effect is particularly noticeable in very noisy conditions. In addition we find that to have maximum noise robustness, the bandpass filters should focus on low modulation frequencies. This reenforces our intuition that noise robustness can be increased by exploiting redundancy in those frequency channels which have long enough integration time not to suffer from envelope modulations that are solely due to noise. The ASR system we design based on these findings behaves more similar to human recognition of noisified digit strings than conventional ASR systems. Thanks to the relation between the modulation filterbank and procedures

for computing dynamic acoustic features in conventional ASR systems, the finding can be used for improving the frontends in those systems.

4.1 Introduction

During the last decades a substantial body of neurophysiological and behavioural knowledge about the human auditory system has been accumulated. Psychoacoustic research has provided detailed information about the frequency and time resolution capabilities of the human auditory system (e.g. Fletcher, 1940; Zwicker et al., 1957; Kay and Matthews, 1972; Bacon and Viemeister, 1985; Houtgast, 1989; Houtgast and Steeneken, 1985; Drullman et al., 1994; Dau et al., 997a,b; Ewert and Dau, 2000; Chi et al., 2005; Moore, 2008; Jørgensen and Dau, 2011; Jørgensen et al., 2013). It is now generally assumed that the rate with which the tonotopic representations in the cochlea change over time, the so-called modulation frequencies, is a crucial aspect of the intelligibility of speech signals. Drullman et al. (1994) showed that modulation frequencies between 4 Hz and 16 Hz carry the bulk of the information in speech signals. Modulation frequencies around 4 Hz roughly correspond to the number of syllables per second in normal speech; the highest modulation frequencies are most likely related to changes induced by transitions between phones.[1] Despite the fact that several attempts have been made to integrate the concept of modulation frequencies in automatic speech recognition (ASR) (e.g., Hermansky, 1997; Kanedera et al., 1998, 1999; Hermansky, 2011; Schädler et al., 2012; Moritz et al., 2011), these investigations have not led to the crucial break-through in noise-robust ASR that was hoped for. The performance gap between human speech recognition (HSR) and ASR is still large, especially for speech corrupted by noise (e.g. Lippmann, 1996; Sroka and Braida, 2005; Meyer et al., 2011; Meyer, 2013).

For meaningful connected speech, part of the advantage of humans is evidently due to semantic predictability, but also in tasks where there is no semantic advantage, such as in recognizing digit sequences (Meyer, 2013) or phonemes (Meyer

[1] Brainstem research indicates that the human brain has access to modulation frequencies up to at least 250 Hz. Such modulation frequencies might allow resolving the fundamental frequency of voiced speech, which would provide interesting perspectives for understanding speech in –for instance– multi-speaker environments. However, we limit ourselves to the modulation frequency range that pertains to articulatory induced changes in the spectrum.

et al., 2011), humans tend to outperform machines substantially. Therefore, it must be assumed that acoustic details that are important in human processing are lost in feature extraction or in the computation of posterior probabilities in ASR systems.

There is convincing evidence that some information is lost if (noisy) speech signals are merely represented as sequences of spectral envelopes. Demuynck et al. (2004) showed that it is possible to reconstruct intelligible speech from a sequence of MFCC vectors, but when Meyer et al. (2011) investigated the recognition accuracy of re-synthesized speech in noise by human listeners, they found that in order to achieve the same phoneme recognition accuracy as with the original speech, the re-synthesized speech required a signal-to-noise ratio (SNR) that was 10 dB higher (3.8 dB versus -6.2 dB).

In Macho et al. (2002) it was shown that an advanced frontend that implements a dynamic noise reduction prior to the computation of MFCC features reduces the word error rate. Meyer (2013) showed that advanced features, such as power-normalized cepstral coefficients (PNCC) (Kim and Stern, 2009) and Gabor filter features (Schädler et al., 2012) improve recognition accuracy compared to default MFCCs. The advanced frontend, the PNCC and the Gabor filter features introduce characteristics of the temporal dynamics of the speech signals that go beyond static coefficients enriched by adding deltas and delta-deltas. Therefore, it is quite likely that both HSR and ASR suffer from the fact that a conventional frontend that samples the spectral envelope at a rate of 100 times per second and then adds first and second order time derivatives yields an impoverished representation of crucial information about the dynamic changes in noisy speech.

The research reported here is part of a long-term enterprise aimed at understanding human speech comprehension by means of a computational model that is in conformity with the (neuro)physiological knowledge. For that purpose we want to build a simulation that not only makes equally few, but also the same kind of recognition errors as humans in tasks that do not involve elusive semantic processing. As a first step in that direction we investigate the performance of ASR systems with frontends inspired by an auditory model that has proved to predict intelligibility quite accurately in conditions with additive stationary noise, reverberation, and nonlinear processing with spectral subtraction (Elhilali et al., 2003; Jørgensen and Dau, 2011; Jørgensen et al., 2013; Jørgensen and Dau, 2014).

In addition, we investigate how an exemplar-based procedure for estimating the posterior probabilities of sub-word units interacts with the auditory-based frontends.

Auditory models predict speech intelligibility on the basis of difference between the long-term average power of the noise and the speech signals at the output of the peripheral auditory system (Jørgensen and Dau, 2011). However, it is evident that the long-term power spectrum of a speech signal is not sufficient for speech recognition. Auditory models are silent about all the processing of their outputs that is necessary to accomplish speech recognition. As a consequence, it is not clear whether an auditory model that performs well in predicting intelligibility for humans based on the noise envelope power ratio, such as the SNR_{env} model (Jørgensen and Dau, 2011) is also optimal in an ASR system that most probably processes the output of the auditory model in a different way than humans do.

The modulation filterbank in the auditory frontend proposed in (e.g. Jørgensen et al., 2013; Jørgensen and Dau, 2014, 2011) consists of a lowpass filter (LPF) and a number of bandpass filters (BPFs) that together cover the modulation frequency band up to 20 Hz. In our work we will vary the cut-off frequency of the LPF, as well as the number and centre frequencies of the BPFs. In this respect, our experiments are somewhat similar to the experiments reported in Moritz et al. (2011), who aimed to harness knowledge about the human auditory system to improve the conventional procedure for enriching MFCCs with delta and delta-delta coefficients. In our research the focus is on understanding how and why resolving specific details in the modulation spectrum improves recognition performance, rather than on obtaining the highest possible recognition accuracy. The way in which we use sparse coding for estimating the likelihood of sub-word units in noise-corrupted speech is very different from the approach pioneered by Gemmeke et al. (2011b), who tried to capture the articulatory continuity in speech by using exemplars that spanned 300 ms. In Ahmadi et al. (2014) it was shown that single-frame samples of the output of a modulation filterbank capture a comparable amount of information about articulatory continuity. In that paper we designed the modulation filterbank based on knowledge collected from relevant literature on the impact of different modulation bands on clean speech recognition. Here, we extend that work substantially by experimenting with conceptually motivated designs of the filterbank.

All theories of human speech comprehension (e.g. Cutler, 2012) and all extant ASR systems (e.g. Rabiner and Juang, 1993; Huang et al., 2001; Holmes and Holmes, 2001) assume that speech recognition hinges on recognizing words in some lexicon, and that these words are represented in the form of a limited number of sub-word units. The recognition after the frontend is assumed to comprise two additional processes, viz. estimating the likelihoods of sub-word units and finding the sequence of words that is most likely given the sub-word unit likelihoods. Both computational models of HSR (e.g. ten Bosch et al., 2013, 2015) and ASR prefer statistical models, or -alternatively- neural network models, for estimating sub-word model likelihoods and some sort of finite state transducer for finding the best path through the sub-word unit lattice.

Despite the analogy between artificial neural networks and the operation of the brain, and despite the fact that networks of spiking neurons have been shown to be able to approximate arbitrary statistical distributions (e.g. Buesing et al., 2011), there is no empirical evidence in support of a claim that human speech processing makes use of statistical models of sub-word units. Therefore, we decided to explore the possibility that the estimation of the likelihoods of sub-word units is mediated by an exemplar-based procedure (Goldinger, 1998). Exemplar-based procedures offer several benefits, compared to GMM-based approaches. An advantage that is especially beneficial for our work is that exemplar-based approaches can handle high-dimensional feature vectors, without the need for dimensionality reduction procedures that are likely to mix up tonotopic features that are clean and features that are corrupted by some kind of 'noise'. In addition, exemplar-based representations are compatible with recent findings about the representation of auditory patterns in human cortex Mesgarani et al. (2014a,b) and models of memory formation and retrieval (e.g. Wei et al., 2012; Meyer, 2013).

De Wachter et al. (2007) have shown that an exemplar-based approach to automatic speech recognition is feasible when using MFCCs and GMMs. More recently, Gemmeke et al. (2011b) and Ahmadi et al. (2014) have shown that noise-robust ASR systems can be built using exemplar-based procedures in combination with sparse coding (e.g. Lee and Seung, 1999; Olshausen and Field, 2004; Ness et al., 2012). Geiger et al. (2013) have shown that the exemplar-based SC approach can be extended to handle medium-vocabulary noise-robust ASR. In sparse coding procedures a -possibly very large- dictionary of exemplars of speech and noise is

used to represent unknown incoming observations as a sparse sum of the exemplars in the dictionary.

The seminal research in Bell Labs by Fletcher (1940, 1953) provides evidence for the hypothesis that speech processing relies on matching incoming signals to stored knowledge in separate frequency bands. That insight has been explored for the purpose of noise-robust ASR in the form of multi-stream processing (Misra, 2006). We apply the same insight to the frequency bands in the modulation spectrum: we assume that the high-dimensional modulation spectrum contains enough features that are not affected by the noise, so that they will dominate the the distance measure in a sparse coding engine. The probability that 'clean' bands exist will depend on the design details of the modulation filter (and on the noise characteristics).

A sparse coding engine that represents noisy speech in the form of sparse sums of clean speech and pure noise exemplars can operate in three main ways. If it starts with matching noise exemplars, the operation is reminiscent of noise suppression and spectral subtraction (e.g. Kolossa and Haeb-Umbach, 2011). If the engine starts with matching speech exemplars, its operation is reminiscent of missing data approaches and glimpsing (Cooke, 2006). Combinations of both strategies can also be envisaged. A third possible strategy, and the strategy used in this chapter, is treating the noise and speech exemplars in the exact same way, leaving it to the solver whether an unknown exemplar is first matched to speech or noise exemplars.

To maximize the possibility for comparing our results to previous research, we develop our initial system using the AURORA-2 data set. Although one might argue that the AURORA-2 task is not representative for a general speech recognition task, the task does not limit the generalizability of the insight gained. Actually, the design of AURORA-2 is beneficial for our current purpose for two reasons. First, recognizing connected digit strings does not require an advanced language model; the fact that all sequences of two digits are equally probable minimizes the interference between the frontend and the backend. This set-up also corresponds to research on human speech intelligibility, which is often based on short semantically unpredictable (and therefore effectively meaningless) utterances. Second, the literature contains a number of benchmarks to which the current results can be compared. In our experiments we will follow the conventional approach to the

FIGURE 4.1: Block diagram of the noise-robust ASR system.

AURORA-2 task which requires estimating the posterior probabilities of 176 speech and 3 silence states in a hidden Markov model.

4.2 System overview

The recognition system used in this work is depicted schematically in Figure 4.1. We discern three main processing blocks. In the first block, acoustic features are extracted every 10 ms from the speech signal using the same type of signal processing as employed in the speech-based envelope power spectrum model (sEPSM) proposed by Jørgensen and Dau (2011) and Jørgensen et al. (2013). The sEPSM model contains more simplifying assumptions than the auditory model proposed in Chi et al. (2005), but the models are very similar in spirit. The feature extraction block is described in more detail in Section 4.2.1. The second block uses the outputs of the modulation filters for estimating the posterior probabilities of the 179 sub-word units (HMM-states) in AURORA-2 by means of a sparse coding (SC) approach (Gemmeke et al., 2011b; Ahmadi et al., 2014). This block is explained in detail in Section 4.2.2. Finally, the third block is a conventional Viterbi decoder that finds the most likely word sequence combining prior and posterior probabilities of the 179 model states. This block is described in Section 4.2.3.

4.2.1 Feature extraction

Figure 4.2 shows a diagram of the feature extraction module. An auditory filterbank consisting of 15 gammatone filters is applied to the 8 kHz speech signal $x(t)$ and forms a set of sub-band signals $X_g(t), g = 1, \cdots, 15$. The centre frequencies of the gammatone filters range from $F_1 = 125$ to $F_{15} = 3150$ Hz, distributed along a

log-frequency scale with $1/3^{rd}$ octave spacing. The gammatone filters were implemented in the time domain. The envelope of each gammatone filter output is then calculated as the magnitude of the analytic signal using the Hilbert transform:

$$\vec{E}_g(t) = |\vec{X}_g(t) + j.\text{Hilbert}(\vec{X}_g(t))|. \tag{4.1}$$

The model proposed in Chi et al. (2005) uses 24 filters per octave. However, it is widely agreed (e.g. Moore, 2008) that a $1/3^{rd}$ octave gammatone filterbank captures all detail in auditory signals that are relevant for speech recognition. Therefore, the design of the gammatone filterbank is kept constant in all experiments.

The 15 sub-band envelopes are downsampled to 100 Hz and then fed into a bank of $M + 1$ modulation frequency filters, one lowpass and M bandpass filters. Thus, the output of the modulation filterbank consists of $15 \cdot (M + 1)$-dimensional feature vectors. In section 4.3 we evaluate the impact on recognition performance when the number of modulation bandpass filters and the way in which their centre frequencies are distributed on the frequency axis are varied.

In the modulation filterbank we used a first-order Butterworth lowpass filter (downward slope −6dB/oct) and a set of second-order bandpass filters with quality factor $Q = 1$ (rising and falling slopes of +6 and −6 dB/oct respectively), since a filterbank consisting of $Q = 1$ filters simulated the intelligibility of human listeners best (e.g. Jørgensen et al., 2013; Jørgensen and Dau, 2014, 2011). The modulation filterbanks were also implemented in the time domain.

The operation of the feature extraction module is illustrated in Figure 4.2. The left-hand column shows the operation in the frequency domain. The right-hand column shows two snapshots of the operation in the time domain. The top panel shows the envelope of the output of the gammatone filter with centre frequency $F_g = 315$ Hz for an utterance of the digit string "zero-six". The bottom panel shows the decomposition of this envelope in its modulation frequency components. The all-positive blue curve in the right-hand bottom panel is the output of the low pass filter; the other curves in this panel represent the output of the modulation bandpass filters. The complete output of the modulation filterbank is a set of time signals $E_{m,g}(t)$ which represent the m^{th} modulation frequency component centred at F_m Hz of the g^{th} gammatone sub-band envelope at F_g Hz. The envelopes at the

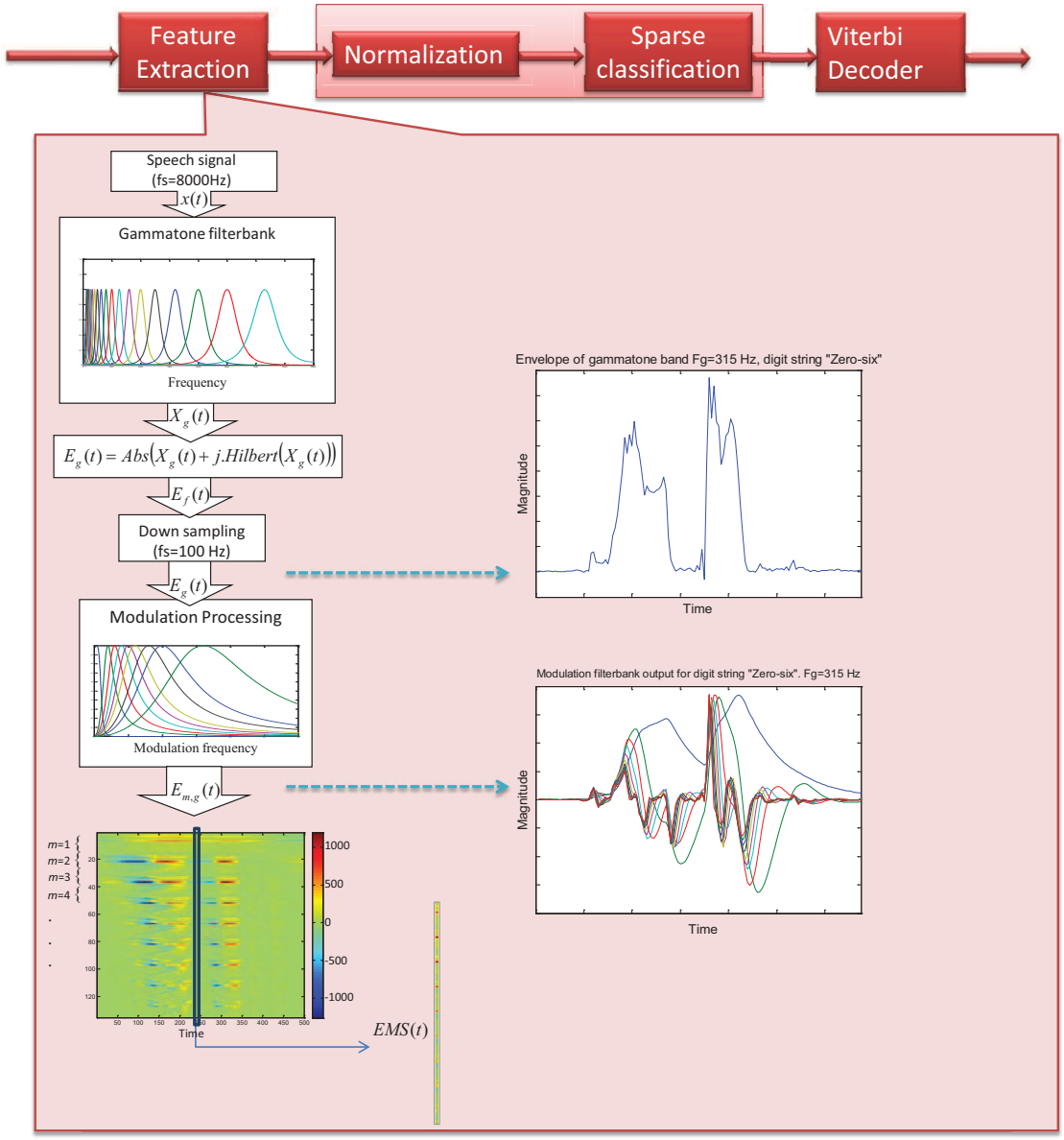


FIGURE 4.2: Block diagram of the feature extraction module. Left column: system operation in frequency domain. Right column: examples of time domain representations.

outputs of the gammatone filters can be approximately reconstructed by means of eq. (4.2).[2]

$$\sum_{m=1}^{M+1} \vec{E}_{m,g}(t) \approx \vec{E}_g(t)\,, \quad g = 1, 2, ..., 15. \tag{4.2}$$

[2]Depending on the spacing of the centre frequencies of the filters, the approximation of eq. (4.2) may be more or less accurate. If a non-uniform resolution over frequency is considered desirable, the resulting sum is a "distorted" version of the original envelope in which the more densely represented frequencies are over-represented/emphasized.

The bottom panel in the left-hand column in Figure 4.2 shows the amplitudes of the outputs of nine modulation frequency filters for each of the 15 gammatone filters for the utterance "zero-six". We refer to this representations as the envelope modulation spectrogram (EMS). The EMS feature vector is obtained by stacking the decomposed sub-band envelopes. Because the signal envelopes are downsampled to 100 Hz, we obtain an EMS feature vector every 10 ms (which we will, analogous to customary ASR terminology, refer to as *feature frames*). Contrary to conventional Mel filter feature extraction, the vector elements do not apply to fixed analysis windows of 25 ms that are shifted with a step size of 10 ms. Instead, the effective time context spanned by the feature value in a modulation band depends on the duration of the impulse response of the corresponding modulation filter. Ahmadi et al. (2014) found that retaining the phase information of the modulation frequency components, i.e., not compensating for the group delay and refraining from applying full-wave rectification to the filter outputs, had a beneficial effect on recognition performance. A similar result was found in Moritz et al. (2011). Therefore, we refrained from reverting to magnitude features and any form of group delay compensation.

4.2.2 Computation of posterior probabilities

The sparse coding procedure needs a dictionary of speech and noise exemplars. In all experiments in this chapter we used a dictionary that comprises 17,148 speech exemplars and 13,504 noise exemplars. For each configuration of the the modulation filterbank a new dictionary was constructed. Exemplars consist of a single feature frame (EMS vector). Given the amplitude response of the modulation filters with the lowest centre frequencies, information about continuity of spectral changes over time is preserved in the EMS features. For all configurations of the modulation filterbank the exact same time frames extracted from the training set in AURORA-2 were used as exemplars.

The speech and noise exemplars were obtained by means of a semi-random selection procedure. We made sure that we had the same number of exemplars from female and male speakers, and almost the same number of exemplars associated with the 179 states in the AURORA-2 task. For that purpose we labelled the clean training speech by means of a conventional HMM system using forced alignment. Most states were represented by 98 exemplars in the dictionary. The

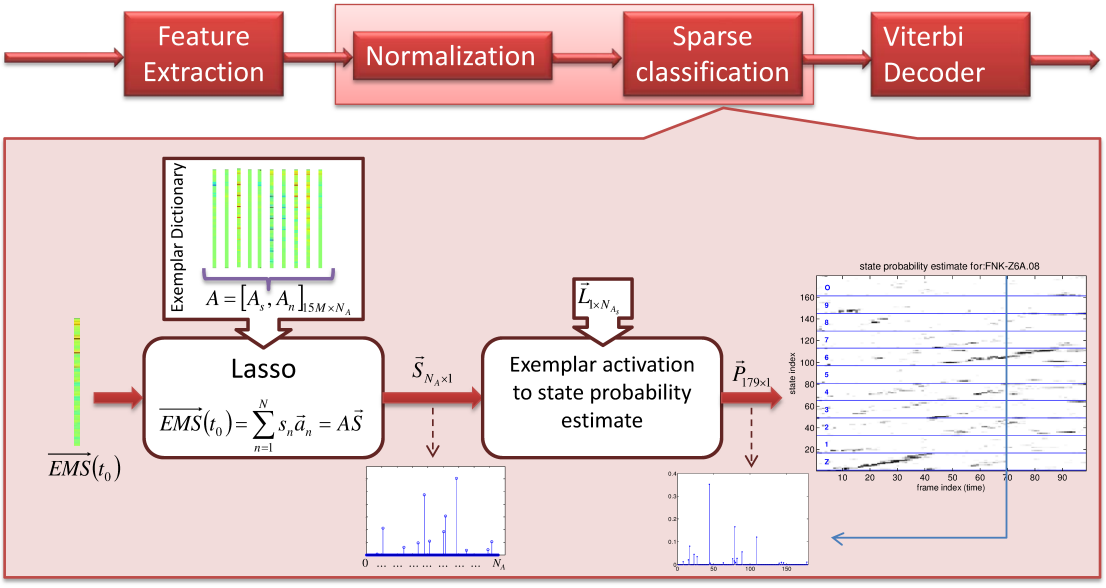


FIGURE 4.3: Block diagram of the posterior probability computation block. A sample posterior probability matrix is visualized in the right side of the figure. The activation vector (S) and state posterior probability vector (P) of a single time frame of the sample signal is shown in the bottom part of the figure.

remaining states, which had fewer frames in the training material, were represented by at least 86 exemplars. To obtain the noise exemplars the noise signals were reconstructed by subtracting the clean speech from the noisified speech in the multi-condition training set. The resulting signals were processed by the modulation frequency frontend, and the noise exemplars were randomly selected from these output signals.

As can be seen in Figure 4.3, the procedure for estimating posterior probabilities of sub-word units consists of several steps. The first step involves a normalization of the EMS features (i.e., standard deviation equalization and Euclidean-normalization), the second implements the reconstruction of unknown observations as a sparse sum of exemplars in a dictionary (sparse coding), and the third step converts the exemplar activations to posterior probabilities.

Standard deviation equalization and Euclidean-normalization We used a Lasso procedure for reconstructing EMS vectors as a sparse sum of exemplars from the dictionary (Efron et al., 2004). Lasso is able to handle the positive and negative components in the EMS vectors. The Lasso procedure minimizes the root mean square of the difference between an observation and its reconstruction.

The range and variance of the components of the EMS vectors differs considerably (Ahmadi et al., 2014). To make sure that all gammatone bands can make an effective contribution to the distance measure, some equalization in the EMS vectors is required. We follow the strategy used in Ahmadi et al. (2014), in which the standard deviations of the samples of the gammatone envelope signals $E_g(t)$ within each modulation band are equalized using weights obtained from the speech exemplars in the dictionary. Each $E_{m,g}(t)$ is multiplied by an equalization weight w_g:

$$w_g = 1/\{\frac{1}{M+1}\sum_{m=1}^{M+1}\sigma_{15\cdot(m-1)+g}\} \qquad for \;\; 1 \le g \le 15, \tag{4.3}$$

where σ_i ($i = 15\cdot(m-1)+g$), $1 \le i \le 15\cdot(M+1)$, is the standard deviation of the ith element of the speech dictionary exemplars. With this procedure the standard deviation of these modified features is equalized within each modulation band, while the relative importance of the different modulation bands is retained. The equalization weights were recomputed for each configuration of the modulation filterbank.

Algorithms for finding the optimal representation of unknown observations in the form of a sparse sum of exemplars are sensitive to the (Euclidean) norm of the observations and exemplars. Therefore, we normalized all exemplars and all unknown feature vectors to unit Euclidean norm. However, for speech-silence segmentation, information about the absolute magnitude of the filter outputs is needed. We used the unnormalized EMS vectors for that purpose.

Sparse coding Unknown observations $\overrightarrow{EMS}(t)$ are reconstructed as a sparse linear combination of exemplars from a dictionary $\boldsymbol{A}$ that contains both speech and noise exemplars,

$$\overrightarrow{EMS}(t) \approx \sum_{n=1}^{N} s_n \vec{a}_n = \boldsymbol{A}\vec{S}, \tag{4.4}$$

where $\vec{S}$ is a sparse weight vector that contains the non-negative exemplar activation scores of the dictionary exemplars that minimize the Euclidean distance

between the test vector $\overrightarrow{EMS}(t)$ and the reconstructed version, subject to a sparsity constraint (controlled by λ):

$$\min \left\| \overrightarrow{EMS}(t) - \boldsymbol{A}\vec{S} \right\|_2 \quad s.t. \quad \left\| \vec{S} \right\|_1 < \lambda. \tag{4.5}$$

From activations to posterior probabilities The exemplar activation scores must be converted into state posterior probabilities. For that purpose, we use the state labels of the speech exemplars in the dictionary. As the exemplar dictionary $\boldsymbol{A} = [\boldsymbol{A}_s, \boldsymbol{A}_n]$ is the concatenation of a noise and a speech dictionary, the activation vector $\vec{S}$ in eq. (4.5) can be split into two separate parts $\vec{S} = \begin{bmatrix} \vec{S}_s \\ \vec{S}_n \end{bmatrix}$, indicating the weights corresponding to speech and noise exemplars, respectively. Since the noise exemplar activations are irrelevant for estimating the posterior state probabilities, we ignore the noise exemplar activations ($\vec{S}_n$). With $\vec{L}_{1\times N_{A_s}}$ the label vector ($N_{A_s} = 17,148$ is the number of speech exemplars), and the i^{th} element $1 \leq L_i \leq 179$ representing the label of the i^{th} exemplar in the speech dictionary, we compute a cumulative state activation vector $\vec{C}$ in which each element $C_j,\ j = 1, 2, ..., 179$ is the sum of the activation scores corresponding to dictionary exemplars that have state label number j:

$$C_j = \sum_{\{i|L_i=j\}} S_i, \tag{4.6}$$

where S_i is the i^{th} element in S_s. The state posterior probability estimate is then computed by normalizing the vector $\vec{C}$ to L_1 norm 1.

$$\vec{P} = \frac{\vec{C}}{\sum_{j=1}^{179} C_j}. \tag{4.7}$$

As in Gemmeke et al. (2011b), it appeared that the procedure of eq. (4.6) systematically underestimates the posterior probability of the three silence states. This is due to the fact that the normalization of all EMS vectors to unit length effectively equalizes the overall magnitude, thereby destroying most of the information that distinguishes silence from speech. Therefore, we implemented an additional procedure that estimates the probability of a frame being either speech or silence

on the basis of the unnormalized feature values. In frames that were classified as silence by this procedure the posterior probability of the three silence states was set to 0.333, and the posterior probability of the 176 speech states was set to some small floor value.

4.2.3 Viterbi decoder

The Viterbi decoder finds the most likely word sequence in a 179 (states) by N (frames) matrix by combining prior and posterior probabilities of the 179 states. The implementation allows us to use different word entrance penalties for the eleven digit words and the silence 'word'. The decoder uses a pre-estimated 179-by-179 state-to-state transition matrix that contains the log probabilities associated to each state-state transition. Probabilities of the non-eligible transitions are first floored to a small positive value before the logarithm is applied. This flooring has a negligible effect on the total probability mass (i.e., the posterior probabilities of the 179 states to which a transition is allowed still sum almost to one). The state-to-state transition matrix is fixed across all experiments in this chapter. The word-word transitions in the language model (LM) are determined by the conditional bigram (word-word) probabilities, which are virtually uniform.

There are two free parameters (i.e. the word and silence entrance penalties) which were tuned on a development test set for adjusting the balance between insertions and deletions and to minimize the word error rate. The decoder only provides the best path with the associated accumulated score and the hypothesized words and silences, including a segmentation at the word level.

4.3 Exploiting modulation frequency domain information

To investigate the impact of the way in which the information about modulation frequencies is represented in the EMS feature vectors, we designed a sequence of experiments. In "Study 1" we use a simplified version of the auditory model to investigate several technical and conceptual issues. We also address the correspondence between the LPF and BPFs in the modulation filterbank on the one hand

and the static and dynamic features in conventional ASR systems (c.f. Moritz et al., 2011). In Study 2 we investigate the performance gain that can be obtained when the cut-off frequency of the LPF is varied and an additional number of modulation bandpass filters are added. Also, we investigate how recognition performance is affected when the LPF and BPFs cover the same modulation frequency range. Finally, in Study 3 we return to the original auditory model (keeping the cut-off frequency of the LPF fixed at 1 Hz), and investigate the impact of different configurations of the bank of BPFs (varying number of BPFs and the spacing of centre frequencies, i.e., linearly or logarithmically) used for capturing the dynamic information.

4.3.1 Study 1: Exploratory experiments

We started experimenting with a highly simplified auditory-like model that consists of a LPF in combination with one BPF that emphasizes modulations in a specific frequency band, i.e., $\boldsymbol{M}$, the number of BPFs in the modulation filterbank equal to one. One conceptual issue concerns the cut-off frequency of the LPF. Different instantiations of the auditory model used quite different LPFs. For example, Moritz et al. (2011) started from the system described in Dau et al. (997a), where the LPF has a cut-off frequency of 6 Hz. This corresponds to an integration time of approximately 170 ms, compared to the 1000 ms integration time of the LPF with a cut-off frequency of 1 Hz in Jørgensen and Dau (2014) that is used here. One might wonder whether such a long integration time can at all be used in experiments with isolated utterances that may have a duration between 0.5 and 3 s. We address the cut-off frequency of the LPF in this study and investigate it further in the next study in configurations with multiple BPFs. In our simplified model, we followed two different strategies in defining the LPF cut-off frequency: 1) the LPF cut-off frequency is fixed at 1 Hz, while the centre frequency of the BPF increases; 2) the LPF cut-off frequency is always 1 Hz below the centre frequency of the BPF, the centre frequency of which increases. We compare the performance of these simplified models with a single LPF covering the same modulation frequency range to evaluate the advantage of emphasizing specific modulation frequencies using the BPF. The number of feature elements in the simplified auditory model (LPF+BPF) is twice the number of feature elements obtained using a single LPF. Moreover, the shape of effective transfer function of

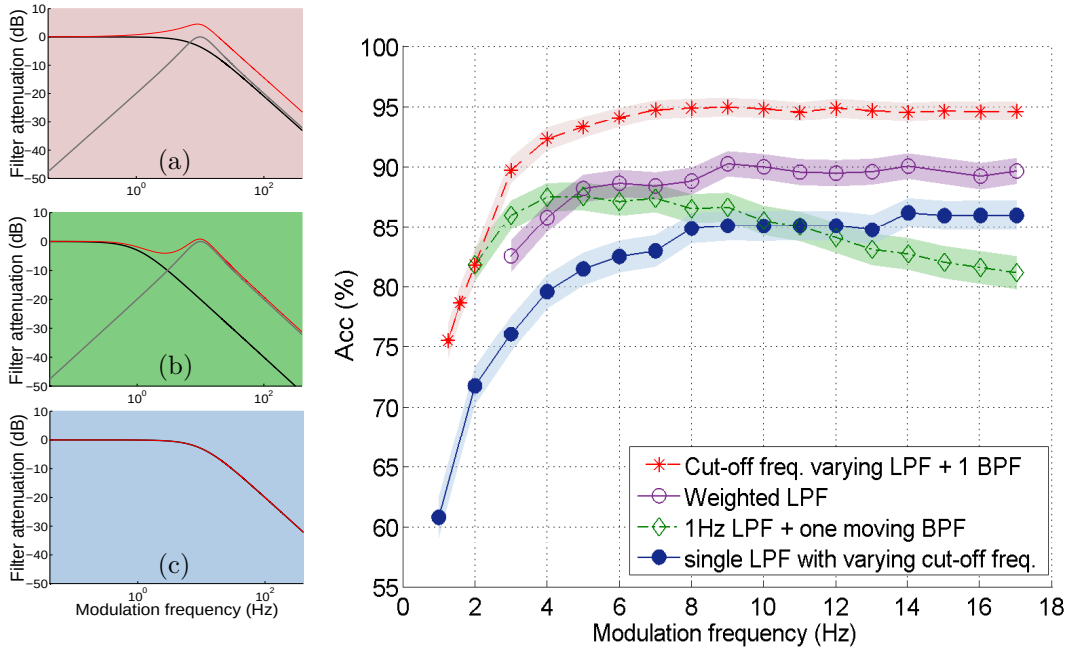


FIGURE 4.4: Word recognition accuracy on clean speech using feature vectors consisting of lowpass filtered gammatone filter envelopes without (blue) or with additional emphasis on a specific modulation frequency band. Emphasis is accomplished by modifying the frequency response of a single lowpass filter (magenta open circles) or by adding an additional bandpass filter. The green curve (diamonds) pertains to a fixed lowpass filter (F_{LP} = 1 Hz) in combination with a bandpass filter with varying centre frequency; the red curve (asterisks) pertains to a lowpass filter of which the cut-off frequency was 1 Hz below the centre frequency of the accompanying bandpass filter. The shaded bands indicate the 95% confidence interval. Sub-figures (a), (b) and (c) show the transfer functions of the composing filters and their sum (in red).

a filterbank consisting a LPF and one BPF is different from a single Butterworth LPF, as shown in Figures 4.4a - 4.4c. To disentangle the effect of these two factors on the performance and also to verify that the effective transfer function is an important issue to consider in the design of a modulation filterbank, we compare the accuracy that can be obtained with a two-filter system and a system with a single LPF that has the same transfer function as the two-filter system. A final, also somewhat conceptual issue that we wanted to explore is to what extent results obtained with a specific configuration for clean speech generalize to the noisified test utterances.

4.3.1.1 Clean speech

The results of the pilot experiments on clean speech are summarized in Figure 4.4. The red curve (asterisks) in Figure 4.4-d shows the recognition accuracy obtained with a modulation filterbank that consists of a LPF with a cut-off frequency that increases from 1 Hz to 16 Hz, combined with a BPF centre frequency 1 Hz higher than the LPF cut-off frequency. Accuracy increases with an increase of the modulation frequency band that is covered, up to a frequency of 7 Hz, where ceiling performance is reached. Interestingly, this 'optimum' is obtained with the cut-off frequency of the LPF in the auditory model proposed in Dau et al. (997a). With 15 gammatone and two filters in the modulation filterbank the EMS feature vectors contained 30 coefficients.

The purple (open circles) curve in Figure 4.4-d pertains to a modulation 'filterbank' that consisted of a single LPF with a frequency response identical to the two-filter system underlying the red (asterisk) curve. Since the modulation filterbank comprised only a single filter, the EMS vectors contained 15 features. From this comparison it can be concluded that representing an overall frequency response by means of two filters, resulting in EMS vectors that contain two sets of 15 features is advantageous.

The blue (filled circles) curve shows the recognition accuracy obtained with a single LPF with increasing cut-off frequency, and a frequency response that was flat in the pass band. The comparison between this curve and the purple curve shows that an overall frequency response identical to the two-filter system yield better accuracy than a flat response when the EMS vectors contain the same number of features.

The green curve (open diamonds) pertains to the accuracy obtained with a two-filter system in which the cut-off frequency of the LPF was fixed at 1 Hz, while the centre frequency of the BPF was increased from 2 Hz to 17 Hz. For the BPF centre frequency of 2 Hz the configuration is identical to the second configuration in the red (filled asterisks) curve. When the centre frequency of the LPF is 3 Hz it can already be seen that the performance lags, relative to the configuration in which the this BPF is combined with a LPF with a cut-off frequency of 2 Hz (the red curve), despite the equal number of features in the EMS vectors. For centre frequencies of the BPF $>$ 6 Hz the accuracy of the this system decreases with

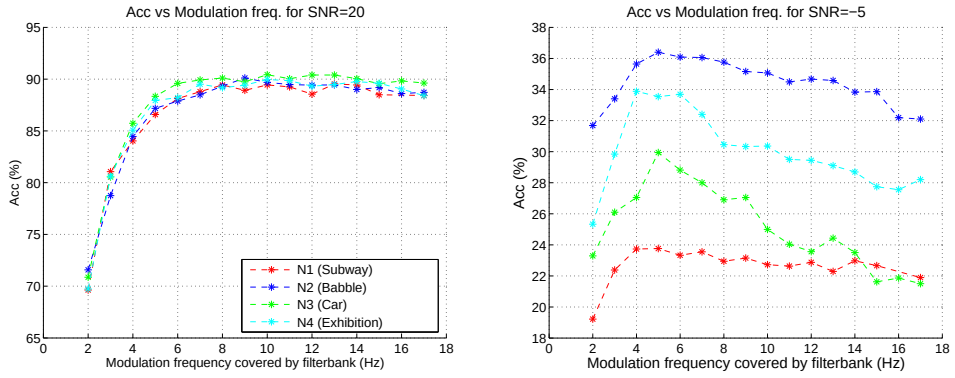
increasing centre frequency. The accuracy of this two-filter system drops below the single LPF system (the purple open circles curve) for BPF centre frequencies > 8 Hz. The accuracy even drops below the single, flat response LPF system for BPF centre frequencies > 14 Hz. We attribute this effect to the overall transfer function of this two-filter filterbank. As can be seen in Figure 4.4-b, the frequency response contains a trough around 4 Hz that deepens as the centre frequency of the BPF increases.

From the data in Figure 4.4 we can draw several preliminary conclusions. Probably the most important conclusions is that the overall frequency response of the modulation filterbank has a large impact on the performance of the system. The frequency response must cover at least the band up to 7 Hz, and emphasizing a somewhat narrow band centred around frequencies up to 7 Hz yield higher accuracy than a flat response. Emphasizing ever higher modulation frequencies has no beneficial, but also no detrimental effect. The second conclusion is that the number of coefficients in the EMS feature vectors is important. With identical frequency responses, the systems that encode the output of the BPF as an additional set of 15 coefficients always perform much better. This indicates that EMS vectors that distribute information about the overall frequency response in a set of features that correspond to the flat part of the shape and the region of the frequency axis that is emphasized are more discriminative.

4.3.1.2 Noisy speech

Since filterbanks that combine a LPF with increasing cut-off frequency with a BPF with centre frequency 1 Hz above the cut-off appeared to yield the best accuracy, we tested these configurations on the noisy utterances of test set A. The other (inferior) configurations mentioned above were also tested; results are not shown, because they do not contribute additional information. Figure 4.5 shows the accuracies in the $SNR = 20$ dB and $SNR = -5$ dB conditions.

From Figure 4.5a it can be seen that the results in the $SNR = 20$ dB condition are similar to the results obtained with clean speech. However, the frequency range at which ceiling performance is reached differs slightly between the four noise types. Also, the extent to which the accuracy varies on the plateau seems to differ slightly between the four noise types.



(a) (b)

FIGURE 4.5: Word recognition accuracy on noisy speech (four noise types in test set A), using feature vectors consisting of a lowpass filtered gammatone filter envelopes together with an additional bandpass filtered version of the envelope. (a) Word recognition accuracy on four noise types at SNR level of 20dB. (b) Word recognition accuracy on four noise types at SNR level of −5dB. (Note the different scales of the vertical axes.)

At the SNR= −5 dB level (c.f., Figure 4.5b), a different pattern of results is visible. Although it is not safe to draw strong conclusions from very low recognition accuracies, several observations stand out. First, there is substantial difference between the noise types. Noise type N2, babble noise, yields the highest accuracies for all LPF cut-off frequencies. The accuracy with car noise (N3) drops almost to the level of subway noise (N1) with cut-off frequencies ≥ 12 Hz. It can also be seen that all four noise types show a decreasing accuracy when the cut-off frequency of the LPF increases beyond some maximum. For car noise the fall is deep and steep, whereas it is quite shallow for subway noise.

An in-depth analysis of the distributions of the EMS vectors showed that these somewhat surprising results are caused by the difference (or similarity) between the two-band EMS features of speech and the corresponding features of the four noise types. In the lower SNR conditions (and especially with SNR= −5 dB) we see two different effects. Noise exemplars in the dictionary account for a substantial proportion of the approximation of the noisy speech EMS vectors; this results in low –and possibly random– activations of the speech states. Except for the subway noise, the reduction of the total activation of speech states becomes worse as the BPF emphasizes higher modulation frequencies, which are less informative for speech. The overall reduction of the activation of speech states is combined with an increasing shift of the activations towards a small number of speech states that happen to have EMS vectors that are somewhat similar to the vectors that

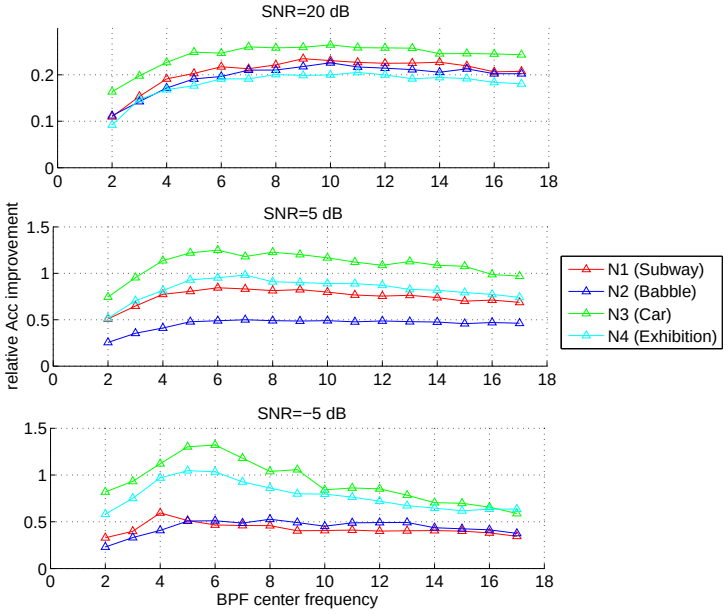


FIGURE 4.6: Relative recognition accuracy (Acc) improvement obtained by adding an additional bandpass filtered version of the envelope to the 16Hz low-pass filtered one. The subplots show the results on three different SNR levels of noisy speech with four different noise types of test set A.

characterize the noises. This results in a digit confusion pattern that strongly favours the digits that happen to contain these favoured states. This effect is especially clear for N1 (subway) and N4 (exhibition hall), whose EMS vectors are characterized by high values in the high-frequency gammatone filters, both in the LPF and BPF. The EMS vectors of N1 show this effect already at low cut-off frequencies, which explains the fairly flat shape of accuracies as a function of cut-off frequency. Babble noise behaves differently in that it does not favour a small number of speech states. The especially detrimental effect of N3 (car) is due to a combination of the two effects: a small number of speech states is favoured, while the total activation of the speech states is small. The large differences between the recognition accuracies with the four noise types at −5 dB SNR suggest –unsurprisingly– that a two-filter modulation filterbank does not provide sufficient resolution for coping with different noise types.

4.3.1.3 The link with delta coefficients in conventional ASR

In addition to commonalities between the acoustic features used in conventional ASR and the output of an auditory model, there are also substantial differences.

The conventional ASR approach is based on (power) spectra estimated from 100 overlapping windows per second. Such a spectrum can be considered as equivalent to the EMS features in a LPF with cut-off frequency set to 50 Hz. Furthermore, conventional delta coefficients in conventional ASR (i.e., the time derivatives of the static features) can be viewed as the output of a single modulation frequency bandpass filter. The transfer function of a differentiator has a rising slope of +6 dB/octave; therefore, the output of a bandpass filter with a rising slope of +6 dB/octave can be considered as a low-pass filtered version of a differentiator. The falling slope of the BPF determines to what extent the high frequencies in the differentiated signal 1are attenuated. In the $Q = 1$ filters of our auditory model, the falling slope is -6 dB/octave. In conventional ASR the centre frequency of the bandpass filter, as well as the steepness of the falling slope, depend on the number of static coefficients involved in the regression function used in computing the deltas. With `DELTAWINDOW=5` and a frame rate of 100 frames per second in HTK (Young et al., 2009) the centre frequency of the 'delta' filter is approximately 7.5 Hz, while the attenuation at the Nyquist frequency of 50 Hz is approximately 20 dB.

To obtain a better understanding of the effect of centring the 'delta' filter at different frequencies, we carried out an experiment in which we combined a 16 Hz cut-off frequency LPF with a single BPF with a centre frequency that varied between 2 Hz and 16 Hz. The recognition accuracies obtained with these configurations were compared to the accuracy obtained with a single LPF with cut-off frequency 16 Hz. Figure 4.6 shows the relative improvement for the four noise types for SNR levels of $20, 5$, and -5 dB. A comparison between the curves for the SNR levels shows that the gain increases as the SNR level decreases: While the relative improvement is of the order of 20% to 25% in the $SNR = 20$ dB condition, the performance is improved by 50% up to 130% (noise type dependent) in the SNR= -5 dB condition. Especially at SNR-5 dB the centre frequency at which the recognition accuracy increases most depends on the noise type. This confirms that a single 'delta' filter is not sufficient for making the EMS features robust against different noise types.

4.3.2 Study 2: Multi-resolution representations of modulation frequencies

It is quite likely that humans pay selective attention to the spectro-temporal input when understanding speech, and that selective attention becomes more important as the listening conditions grow more adverse. The gammatone filters allow for a sufficient degree of selectivity in the frequency domain. The subsequent modulation filterbank must provide the selectivity in the modulation frequency domain. In combination with the sparse coding approach for obtaining the posterior probabilities of the 179 states in the AURORA-2 task, a multi-resolution representation, with its attendant longer feature vectors, might enhance the probability that 'correct' clean speech exemplars in the dictionary have a small Euclidean distance to noisy speech frames, because the energy of the noise is much smaller than the energy of the speech in some regions of the EMS vectors. If this is indeed the case, a multi-resolution representation should enhance the resulting recognition accuracy.

In Section 4.3.1 it was concluded that modulation frequencies in the band up to 16 Hz must be covered and that the largest gain in performance relative to a configuration with a single LPF is obtained by emphasizing different modulation frequencies for different noise types and different SNR levels. Therefore, it can be expected that a configuration in which multiple BPFs separate the modulations in different frequency bands would outperform a configuration that contains only a LPF and a single BPF. Auditory models do precisely this, by combining a LPF with a bank of BPFs. Such a filterbank can be configured in two different ways: the BPFs can cover the frequency range above the cut-off frequency of the LPF, or the frequency ranges of the BPF and LPF may overlap, so that the BPFs provide additional resolution in a band that is already covered. Below, we compare these configurations. By doing so, we address two questions: 1- In which modulation frequency range is a high resolution most beneficial for noisy speech recognition? 2- to what extent is it beneficial to represent modulation frequencies both in terms of static and dynamic features by choosing overlapping LPF and BPFs?

In the first experiment we employed a modulation filterbank consisting of a LPF with a variable cut-off frequency (ranging from 1 to 16 Hz), augmented with a bank of BPFs with centre frequencies (spaced 1 Hz apart) covering the range from 1 Hz above the cut-off frequency of the LPF up to 16 Hz. Obviously, the total number of filters in the filterbank ($M + 1$), and therefore the total number of features in

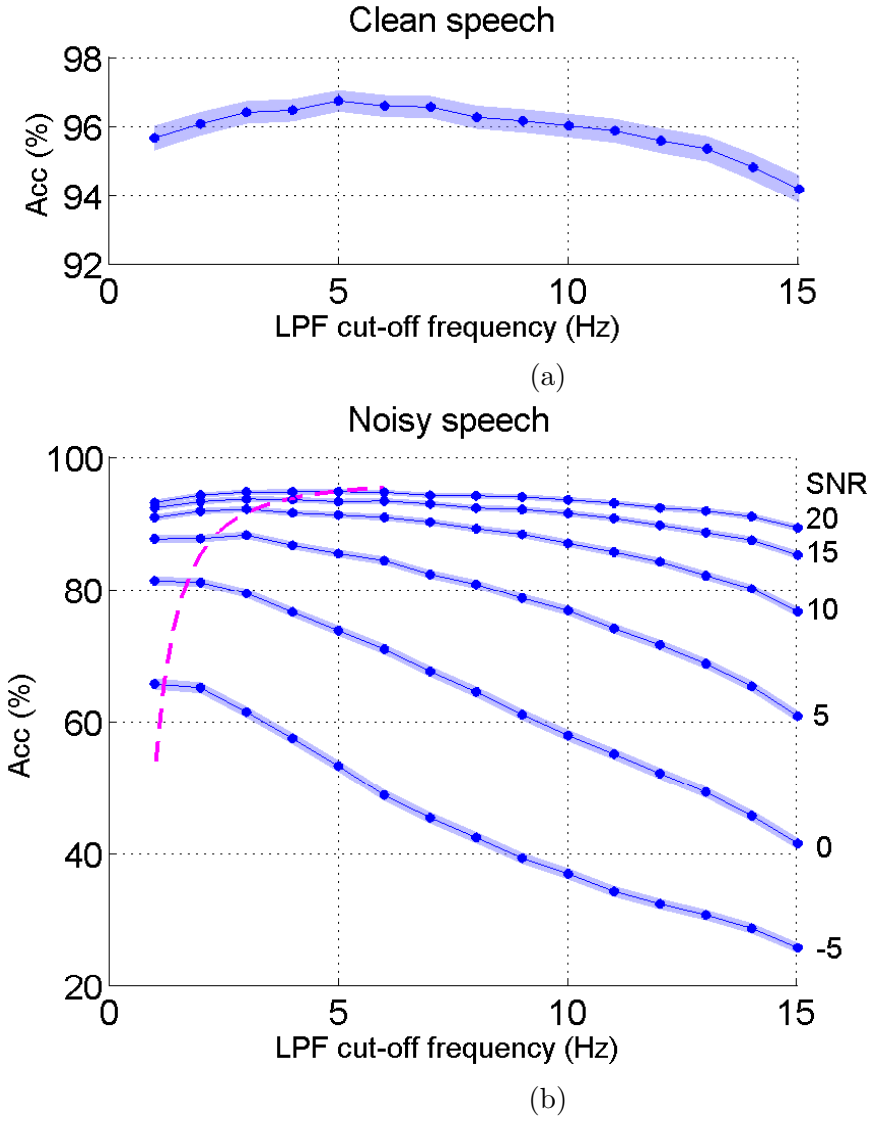


FIGURE 4.7: Word recognition accuracy obtained with feature vectors covering the modulation frequency range of $0 \cdots 16$ Hz. The modulation filterbank consisted of a single lowpass filter with variable cut-off frequency and a variable number of additional bandpass filters with centre frequencies spaced 1 Hz apart to cover the interval beyond the LPF cut-off frequency. Results for clean (top) and noisy speech (bottom) are shown in separate panels to improve resolution. The shaded bands indicate the 95% confidence interval. The dashed line shows the trajectory of the peak position across SNR level.

the EMS vectors ($15 \cdot (M + 1)$), will increase as the cut-off frequency of the LPF decreases. The test is performed on all the clean and noisified data in test set A of AURORA-2. The results of this experiment (averaged over four different noise types) are summarized in Figure 4.7.

The first observation that can be made from the figure is that the configurations

with the largest number of modulation BPFs do not always yield the best recognition accuracy: the curves for the highest SNR levels start with a (small) interval in which the performance is increasing as the cut-off frequency of the LPF increases, corresponding to a decrease of the total number of filters. The cut-off frequency at which the maximum accuracy is obtained is clearly dependent on the SNR level. In the clean condition, the best performance is obtained when the LPF cut-off frequency is 5 Hz and the modulation frequency range of $6 - 16$ Hz is covered by $M = 11$ linearly spaced BPFs. At lower SNR levels, the LPF cut-off frequency at which the maximum accuracy is obtained shifts towards lower frequencies as illustrated by the dashed line in Figure 4.7b: it interpolates the LPF cut-off frequency at which the best performance is obtained at different SNR levels. Moreover, the steeper slopes in the curves corresponding to lower SNR levels indicate that increasing the LPF cut-off frequency, and as a result decreasing the resolution in the lower modulation frequencies, is more harmful in the presence of high noise energy. Apparently, separating modulations in the very low frequency bands, which are not very important for the intelligibility of clean speech, enhances the capability of the sparse coding engine to match noisy speech EMS vectors with 'correct' clean speech exemplars.

In the second experiment we combined 15 BPFs with centre frequencies linearly spaced between 1 Hz and 15 Hz with a LPF the cut-off frequency of which was decreased from 15 Hz to 1 Hz. With lower cut-off frequencies the amount of information about the modulations that can be said to be represented twice (in the BPFs and in the LPF) decreases, but all configurations cover the modulation frequencies up to 16 Hz. Also, the total number of features ($15 \cdot 16$) was identical in all configurations, because the number of filters was fixed.

It appeared that decreasing the cut-off frequency of the LPF from 16 Hz to 1 Hz had no effect on the resulting recognition accuracy. The performance was independent of the cut-off frequency and always equal to the accuracy corresponding to LPF cut-off frequency of 1 Hz in Figure 4.7. From this experiment it can be concluded that the 1 Hz cut-off frequency of the LPF in the model of Jørgensen and Dau (2014) is to be preferred over the 6 Hz cut-off frequency in the model of Dau et al. (997a), especially in low SNR conditions. Apparently, a high resolution in the modulation filterbank is almost always beneficial. The only exception is formed by the conditions with a very high SNR level, where a high resolution in the very low modulation frequencies has a small negative effect.

4.3.3 Study 3: The auditory model revisited

Now that we know that a set of modulation BPFs that cover the frequency range from $2 \cdots 16$ Hz, in combination with a LPF with a cut-off frequency as low as 1 Hz, can yield promising recognition accuracies, we can return to the question whether the 'standard' configuration in auditory models, i.e., $Q = 1$ BPFs spaced at one octave intervals, is the optimal configuration for ASR applications. To address this question we carried out experiments in which the envelopes of the gammatone sub-bands are processed by a number of different modulation filterbanks. The filterbanks consisted of a fixed LPF with a cut-off frequency at 1 Hz and a variable number of BPFs with quality factor $Q = 1$.

4.3.3.1 LPF at 1Hz and BPFs with different distribution patterns

We first compare the recognition performance using filterbanks with similar frequency coverage, but with different number of BPFs and distribution patterns of centre frequencies. The centre frequencies of the BPFs were chosen in three different manners: linearly spaced at 1 Hz distance, logarithmically spaced at $1/3^{\text{rd}}$ octave and at full octave distance. The number of BPFs is gradually increased, adding modulation bands, until they cover the frequency range up to 25 Hz.[3] In Figure 4.8a the recognition accuracies for clean speech of each of these filterbanks are depicted as a function of the centre frequency of the last BPF included (red: linear spacing; purple: octave spacing; green: $1/3^{\text{rd}}$ octave spacing). Note that, as a consequence of the different distribution patterns of the BPFs, the number of BPFs used for covering the range up to a given modulation frequency was different (14 with linear spacing, 4 with octave spacing, 12 with $1/3^{\text{rd}}$ octave spacing, and 10 with the first two filters in the $1/3^{\text{rd}}$ octave spacing left out).

The first observation from this figure is that adding more BPFs improves recognition accuracy, but a ceiling performance is reached when the centre frequency of the last-added filter is 16 Hz. The highest word recognition accuracy is obtained with the linear spacing strategy and amounts to 96.13%, an improvement of approximately 2.2% absolute compared to the best performance obtained with

[3] We increased the modulation frequency range compared to previous experiments. This was done to verify that the logarithmically spaced BPFs (that exhibit a wider spacing at high frequencies) also yielded ceiling performance above approximately 16 Hz.

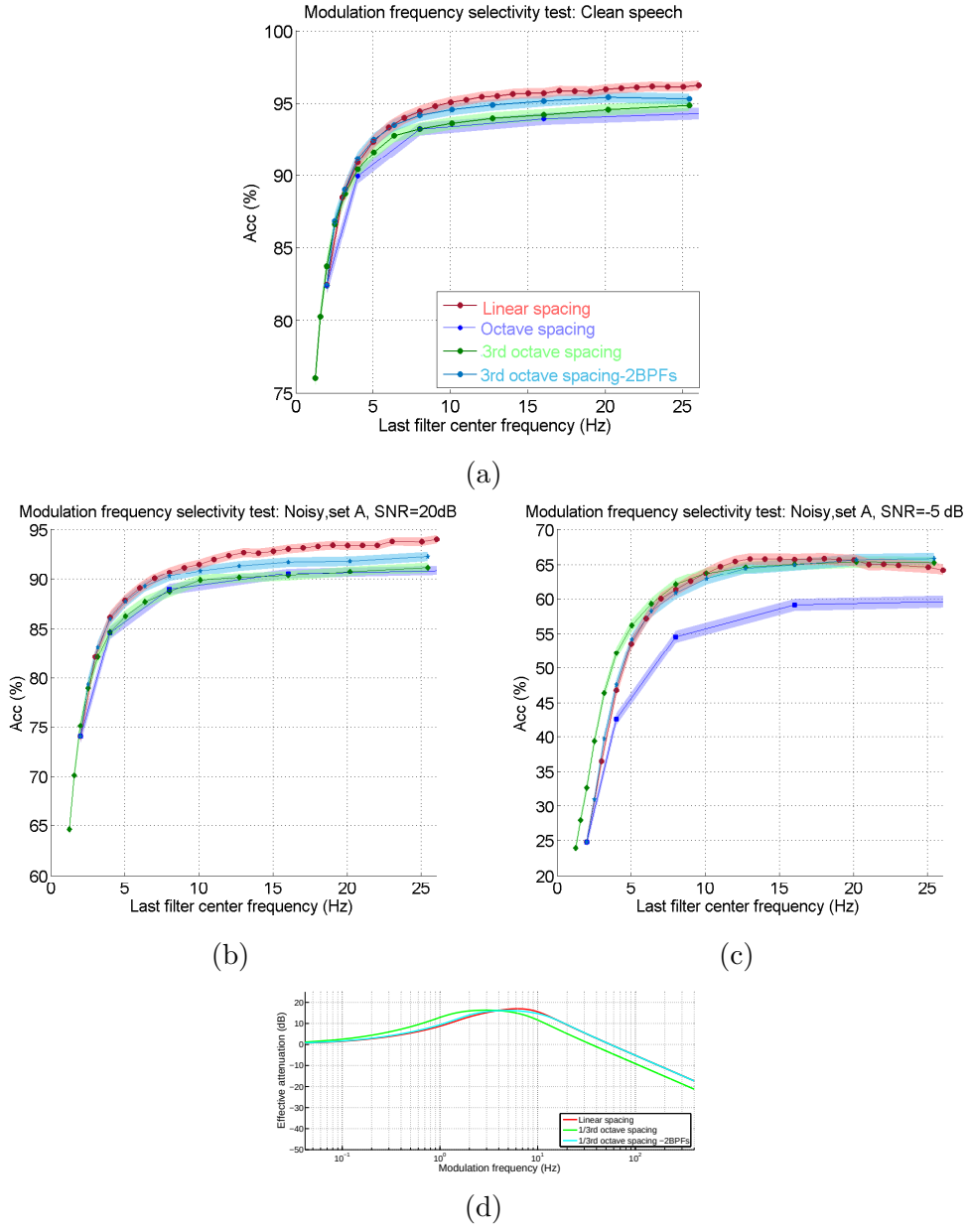


FIGURE 4.8: Word recognition accuracy for (a) clean speech, (b) noisy speech SNR= 20 dB, (c) noisy speech SNR= −5 dB, as a function of the highest centre frequency in the bank of bandpass filters. The shaded areas represent the 95% confidence intervals. The centre frequencies of the filters (F_{BP}) are spaced linearly at 1 Hz intervals (red), or logarithmically at full octave intervals (purple) or at 1/3 octave intervals (green). The blue curve depicts the results obtained with the same 1/3 octave filterbank, without the two filters with $F_{BP} < 2$ Hz. Sub-figure (d) shows the effective transfer functions of the modulation filterbanks with LPF cut-off frequency 1 Hz and 10 (or 8 for the blue curve) BPFs.

a combination of a LPF with cut-off frequency 15 Hz and a single BPF with centre frequency 16 Hz (cf. Study 1).

The second observation from Figure 4.8a is the consistent and statistically significant advantage of the linearly spaced filterbank (the red curve) over the logarithmically spaced filters (the purple and green curves). The difference in number of filters employed cannot explain this observation: to cover the range up to a modulation frequency of 10 Hz, the $1/3^{\text{rd}}$ octave spacing and the linear spacing require ten and nine filters respectively; still, the linearly spaced filterbank outperforms the $1/3^{\text{rd}}$ octave spacing. Also, despite the different number of filters, the octave spaced and the $1/3^{\text{rd}}$ octave spaced filterbank have roughly equal performance. Therefore, the most plausible explanation lies in the fact that different locations of the centre frequencies of a set of BPFs yield different effective transfer functions. To illustrate this effect, we plotted the effective transfer function for the linear, octave and the $1/3^{\text{rd}}$ octave spaced filterbanks with $M = 10$ in Figure 4.8d. Clearly, the $1/3^{\text{rd}}$ octave spaced filterbank emphasizes the very low modulation frequencies much more than the linearly spaced filterbank (peak at 3.0 Hz compared to 6.25 Hz).[4]

From the perspective of sparse coding this means that information about modulations in a frequency range that exhibits non-negligible variance, but contains little information about the contents of speech signals, may have too strong an impact on the Euclidean distance measure, giving rise to sub-optimal recognition performance. To test this hypothesis, we removed the first two BPFs from the $1/3^{\text{rd}}$ octave spaced filterbank ($f_c = 1.26$ Hz and $f_c = 1.58$ Hz). As a result, the effective transfer function of the modified filterbank does no longer over-emphasize the lowest modulation frequencies compared to the linearly spaced filterbank (cf. Figure 4.8d). Consequently, as shown by distance between the green and the light blue curve in Figure 4.8a, the recognition accuracy for clean speech was always higher than the results with the corresponding full $1/3^{\text{rd}}$ filterbanks. Note that the content of the BPFs with $F_c = 1.26$ Hz and $F_c = 1.58$ Hz do contain some useful information since the performance levels at the two left most points in the red curve of Figure 4.4d are larger than the left-most point of the blue curve in Figure 4.4d (using one static feature only). However, in combination with more

[4]The frequency response of the filterbank with octave spacing is not shown, because the centre frequency of a substantial number of 10 filters is beyond the Nyquist frequency.

BPFs covering a larger range of modulation frequencies, a modulation frequency range that is sampled too densely at the low end is harmful for recognition.

In this experiment we compared filterbanks with different numbers of filters. We also created filterbanks with the same number of filters as in the filterbank with linearly spaced BPFs, equally spaced on a logarithmic frequency axis. None of these configurations appeared to provide better performance than the linear spacing of the centre frequencies.

Figures 4.8b and 4.8c show the results obtained with increasing numbers of differently spaced filters for the two extreme noise conditions, i.e., SNR=20 dB and SNR=−5 dB. In the SNR=20 dB condition the superiority of the linear spacing, with a (much) larger number of filters, is more apparent than in the clean speech condition. In the SNR=−5 dB condition the filterbanks with octave spacing, and therefore smaller numbers of filters, yield much lower accuracies than the configurations with higher numbers of filters. This suggests that, particularly in noisy conditions, the sampling of the modulation frequency domain needs to be sufficiently fine-grained for the ASR system to reap the maximum possible benefit from the multi-resolution representation. It can also be seen that the recognition accuracy obtained with linearly spaced filters starts decreasing when filters with centre frequencies > 16 Hz are added. The modulations in these frequency bands are mainly associated to the noise. This confirms our earlier conclusion that it is counter-productive to dedicate a substantial proportion of the EMS features to modulation frequency bands that do not contain information relevant for speech recognition. From Figure 4.8c it can also be seen that a larger number of BPFs is not always beneficial: the configuration with fewer $1/3^{\text{rd}}$ octave spaced filters is clearly competitive.

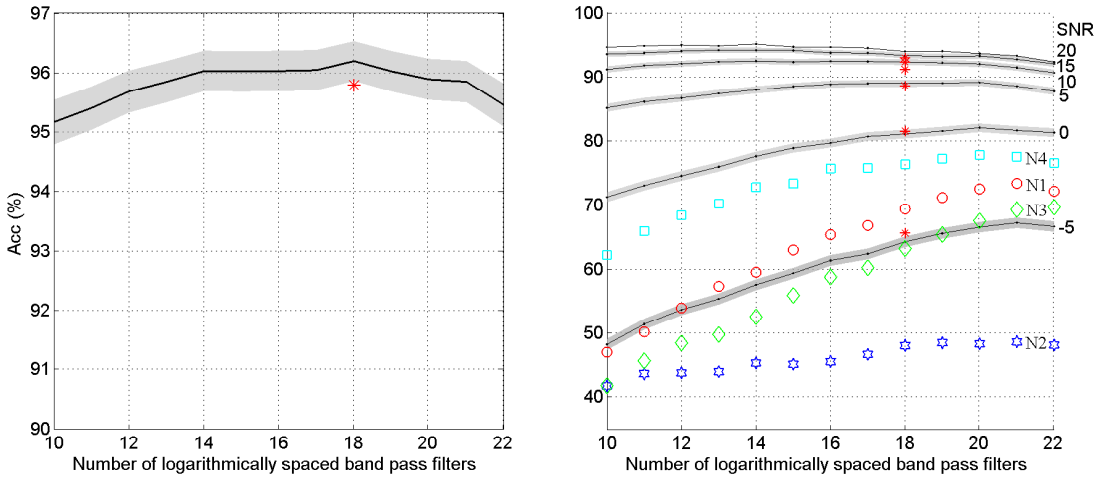
4.3.3.2 LPF at 1Hz and varying number of BPFs logarithmically positioned to approximate a given effective transfer function

In most of the previously described experiments there was an interaction between the total range of modulation frequencies covered, the number of filters in that range and the distribution patterns of the centre frequencies. The fact that leaving out the lowest-frequency filters from the $1/3^{\text{rd}}$ octave filterbank improved the recognition accuracy suggests that the presence of irrelevant features incurs the

risk that the Euclidean distance in the sparse coding process homes in on exemplars that fit these irrelevant features, at the cost of the features that do matter. The shape of the effective transfer function of the filterbanks and the frequency at which the response is maximal indicate which modulation frequencies will be represented with many features and dominate the Euclidean distance measure in the Lasso decoder. In study 1 it was found that the effective transfer function can be used as a criterion for comparing different filterbank configurations. Therefore, we conducted an experiment in which we used the effective transfer function of the best-performing linearly spaced filterbank (i.e. 19 filters: 1 LPF + 18 BPFs) as a target that we try to approximate by means of a variable number of logarithmically spaced BPFs. In contrast to the previous experiments, however, we allowed the centre frequencies of the first and last filter in the filterbank to vary. Imposing the additional condition that the resulting configurations would provide at least some resolution in the low modulation frequency range, without pushing the lowest centre frequency below 2 Hz and without pushing the highest one above 36 Hz (so that the −3 dB point of the falling slope of the BPF does not exceed the Nyquist frequency), we ended up with configurations with a minimum number of 10 and a maximum number of 22 filters. The recognition accuracy results obtained with these configurations are shown in Figure 4.9.

As can be seen from Figure 4.9a, close to maximum performance on clean speech can be achieved with any number of log-spaced BPFs with $M \geq 15$. The best performance is achieved with 18 BPFs; the centre frequency of the first and last BPFs are 3.26 Hz and 20.24 Hz, respectively. Although the number of filters is equal to the target linear filterbank, the achieved recognition accuracy is even slightly (but significantly) higher than with the 18 linearly spaced BPFs (0.4% relative; the red asterisk that indicates the accuracy with linearly spaced filters is just beyond the 95% confidence interval).

Figure 4.9b shows the corresponding results for the noisy test utterances from set A at SNRs ranging from 20 dB down to −5 dB. For the highest SNR conditions the accuracy does not improve substantially when the number of filters is increased from 10 to 18. For the lowest three SNR levels increasing the number of BPFs does improve accuracy. In all cases a larger number of filters results in a higher resolution in the lowest modulation frequencies. For SNR=-5 dB, using $M = 21$ BPFs (rather than $M = 18$) yielded a 5% relative improvement. In this configuration the centre frequencies of the lowest and highest BPF were 2.25 Hz



(a) (b)

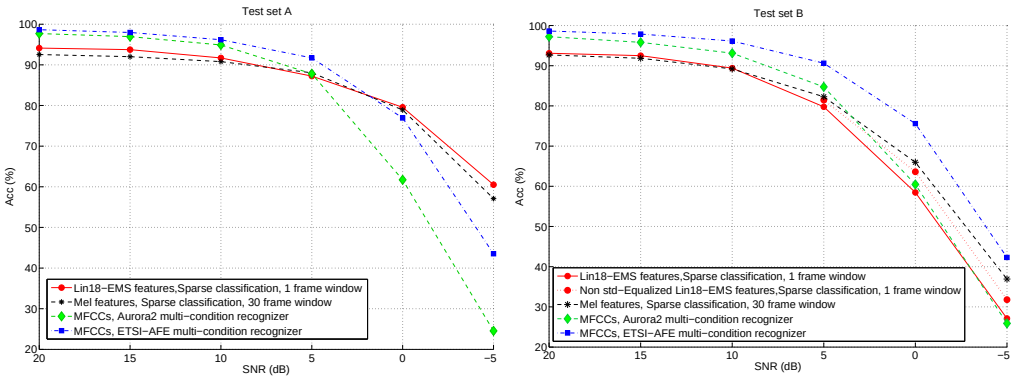
FIGURE 4.9: Word recognition accuracy for (a) clean speech and (b) noisy speech as a function of the number of logarithmically spaced filters. The black lines show the recognition accuracy averaged over all the four noise types (The shaded areas represent the 95% confidence intervals). At SNR=-5, the individual noise results are also plotted in scattered markers: red circle:N1 (Subway), blue hexagram: N2 (Babble), green diamond: N3 (Car) and cyan square: N4 (Exhibition). The red asterisks shows the recognition accuracy obtained with the best performing linear filterbank (1LPF+18 BPFs).

and 19.3 Hz. Figure 4.9b breaks out the recognition accuracies obtained with the four noise types in the $SNR = -5$ conditions. Increasing the resolution of the modulation filterbank has the smallest effect for the babble noise. This was to be expected, because it is unlikely that there are many modulation frequency bands in which babble noise differs substantially from speech.

4.4 Comparison with other ASR systems and HSR

4.4.1 ASR

In this research we investigated how different configurations of the modulation filterbank affect recognition performance. To deepen our understanding of the



(a) (b)

FIGURE 4.10: Word recognition accuracy per test set as a function of ASR for four different systems. 1- The proposed EMS features (Lin18-EMS). 2- Sparse classification results using Mel-spectra features (Gemmeke et al., 2011b). 3- Aurora2 multi-condition recognizer applied to MFCC features (Hirsch and Pearce, 2000). 4- ETSI-AFE multi-condition recognizer applied to MFCC features (Hirsch and Pearce, 2006).

strengths and weaknesses of the combination of EMS features and SC, we compared the performance on test sets A and B in AURORA-2 with previously published recognition accuracies of three other systems: the 'standard' AURORA-2 system trained with the the multi-condition data (Hirsch and Pearce, 2000), the multi-condition AURORA-2 system that includes the Wiener filter based ETSI advanced frontend (Hirsch and Pearce, 2006), and the SC-based system of Gemmeke et al. (2011b). The first two systems use GMMs based on MFCC features to estimate state posterior probabilities, while the third one used Mel-frequency energy spectra as stacks of up to 30 frames and used non-negative matrix factorization with the Kullback-Leibler divergence as the solver in the sparse coding engine. Since there is no configuration of the modulation filterbank that is optimal for all SNR levels and all noise types, we conducted the comparison with the modulation filterbank consisting of the 1 Hz cut-off frequency LPF and $M = 18$ linearly spaced BPFs (which we refer to as the Lin18-EMS system). The Lin18-EMS system is a good compromise between the highest-possible performance for clean speech and the conditions with the lowest SNR level. The detailed results obtained with the Lin18-EMS system are collected in Table 4.1. .

In Figure 4.10, the recognition accuracies of the Lin18-EMS system and the three competing systems is plotted. Figure 4.10a shows the test results for matched noise types in test set A. While the Lin18-EMS system outperforms both MFCC-based multi-condition recognizers at very low SNR levels, its performance at higher

SNRs is substantially worse than the MFCC-based systems. The single-frame EMS features almost always outperform the 30-frame Mel features.

However, the results of the Lin18-EMS system on test set B, pertaining to the unseen noise type conditions, shown in Figure 4.10b, show that our system does not generalize well to unseen noise types, a characteristic that it shares with the other exemplar-based system. The superior performance of the 30-frame Mel features is most probably due to the fact that Gemmeke et al. (2011b) included artificially constructed noise exemplars that accounted to some extent for the mismatch between the noise exemplars from test set A and the different noise types in test set B. Our EMS-based system did not include artificially constructed exemplars. In cleaner conditions (down to 10 dB) the EMS-based system has roughly equal performance as the other exemplar based system. In contrast to the behaviour for test set A, however, the performance drop in SNRs $< 10\,\mathrm{dB}$ is much steeper. Averaged over the four noise types of test set B, the recognition accuracy is approximately equal to that of the multi-condition trained GMM system without noise reduction.

A detailed analysis revealed that the performance of the Lin18-EMS system in fact *is* very similar to the system of Gemmeke et al. (2011b), except for train station noise (cf. Table 4.1). In search for the cause of this deviant behaviour, we found that omitting the standard deviation equalization step ((4.3) in Section 4.2.2) substantially improved recognition performance for utterances corrupted with train station noise at low SNR levels. This is illustrated by the dotted line in Figure 4.10b, which shows the average performance on test set B (SNR= $5, 0, -5\,\mathrm{dB}$) when excluding the standard deviation equalization for train station noise. Recall that the main purpose of the standard deviation equalization procedure was to equalize the contribution of all gammatone frequency bands. The equalization weight vector was designed -using the speech exemplars from the dictionary- such that the standard deviation of the coefficients in the EMS vector are *o*n average equal in all 15 gammatone filters, without changing the relative magnitude of the coefficients pertaining to the modulation bands. It appeared that the equalization procedure works well for noisified speech, as long as the envelope of the 15 gammatone coefficients in the modulation bands does not change between bands with low and high modulation frequencies. As long as that is the case, applying a fixed equalization vector will not change the average modulation spectrum of the noises. However, there are two noise types that violate this assumption, viz.

	SNR	Clean	20	15	10	5	0	-5	Average
Test A	Subway	94.14	94.84	94.38	88.89	86.98	81.98	66.38	86.80
	Babble	93.62	93.44	92.93	91.9	87.07	73.52	42.05	82.16
	Car	93.56	92.69	92.45	91.65	88.58	80.11	59.68	85.53
	Exhibition	93.55	95.53	95.19	94.38	85.71	82.63	73.93	88.70
	Average	**93.72**	**94.12**	**93.74**	**91.70**	**87.24**	**79.56**	**60.51**	**85.79**
Test B	Restaurant	94.14	89.39	91.56	90.97	84.86	69.11	36.94	79.56
	Street	93.62	93.68	92.53	90.60	83.92	63.27	29.53	78.16
	Airport	93.56	94.87	94.15	91.02	82.58	63.05	28.78	78.28
	Train station	93.55	94.54	91.76	85.00	67.76	38.41	13.24	69.18
	Average	**93.72**	**93.10**	**92.50**	**89.40**	**79.78**	**58.46**	**27.12**	**76.29**

TABLE 4.1: The word recognition accuracy obtained using Lin18-EMS features on AURORA-2 test sets. (For explanation see text)

car noise in test set A and train station noise in test set B. The detrimental effect of the violations in car noise are limited, because it is represented in the noise dictionary exemplars taken from the car noise signals. For the train station noise this is not the case. As a result, the match between the modulation spectra of the speech noisified by adding train station noise and the exemplars in the dictionary deteriorates as the SNR level decreases.

4.4.2 Comparison with HSR

To evaluate the combination of EMS features and sparse coding in terms of human like performance, we re-use the data about the recognition accuracy of ten human listeners on AURORA-2 utterances in Meyer (2013). Meyer used three different criteria: speech reception threshold (SRT), the effect of noise types and the effect of string lengths. SRT is the SNR at which listeners achieve a 50% accuracy; usually it corresponds to the SNR at which the accuracy as a function of SNR has the largest negative slope. The SRT estimated for HSR in Meyer (2013) is around −10.2 dB while for the AURORA-2 system trained with the multi-condition data (Hirsch and Pearce, 2000) the SRT is −1.5dB. From Figure 4.10a, it can be inferred that the SRT of the EMS-based system is well below −5 dB; although it is dangerous to extrapolate the curves, it is reasonable to assume that the SRT for the two exemplar-based systems is close to the human SRT. As can be seen from Figure 4.10b, which represents the noise mismatch case (test set B), our

EMS-based system does not generalize well to unseen noise types. We will come back to this issue in section 4.5.

According to Meyer (2013), the difficult noises for ASR and HSR are different. At SNR=0 and -5 dB, performance of AURORA-2 system trained with multi-condition data the performance for babble noise is higher than for car noise, while HSR shows higher performance for car than for babble noise. From Table 4.1 it can be seen that our EMS+SC system shows the same trend as the human listeners: accuracy with babble noise is lower than with car noise. The same holds for the comparison of airport and train station noise, provided that we solve the equalization issue.

In the human data there is a small but clear drop in accuracy for the longest digit strings, which is probably due to memorization problems. Our EMS-based system does not show this effect. This was to be expected, because an automatic system is not affected by the need to memorize long strings. Our system also does not show the problems with one-digit utterances reported by Meyer (2013) for the 'standard' AURORA-2 systems with multi-condition training. The raw EMS features that we used for speech-silence segmentation yield quite accurate results. Only in a very small proportion of the utterances the endpoint estimates differed from voice onset and offset determined from the forced alignment by more than 16 frames, the minimum number of frames needed to find –or hallucinate– a digit word.

In summary, it can be concluded that the operation of our EMS-plus-SC system for the estimation of sub-word probabilities mimics human speech recognition on a semantics-free task better than more conventional MFCC-plus-GMM systems.

4.5 General discussion

In this chapter we investigated how different configurations of the modulation filterbank in an auditory frontend affect the degree to which an exemplar-based engine can provide accurate posterior probability estimates of sub-word units when recognizing noise-corrupted speech. The auditory model proposed in Jørgensen and Dau (2014), which consists of a LPF with a cut-off frequency of 1 Hz and nine Q = 1 BPFs with centre frequencies one octave apart, served as the point of departure. For estimating the posterior probabilities of the sub-word units, we

used sparse coding and a large dictionary of semi-randomly selected exemplars. We found that BPFs with centre frequencies one octave apart do not provide sufficient resolution of the modulation frequencies for automatic (and maybe also for human) speech recognition. We conjecture that a filterbank with octave spacing between the modulation filters is able to discover noise conditions that will certainly compromise intelligibility, but that this configuration may not accurately predict specific confusions that would occur in tasks that require participants to distinguish confusable sounds in the absence of semantic predictability.

From our experiments it appears that there is no unique configuration of the modulation filterbank that is optimal for all SNR levels and all noise types. However, it is safe to conclude that a filterbank consisting of a LPF with cut-off frequency 1 Hz and about $M = 18$ BPFs with centre frequencies between 2 Hz and 20 Hz will provide accuracies close to optimal for most conditions. Centre frequencies of the BPFs with equal spacing on a linear or on a logarithmic frequency axis yielded very similar results. In the SNR-5 dB condition the best results were obtained with a configuration that comprised $M = 21$ logarithmically spaced BPFs, with the lowest BPF centred at 2.25 Hz. In all experiments we found that the lowest SNR levels benefited from a large resolution in the lowest modulation frequencies; however, for the highest SNR levels a very high resolution in the modulation frequency band < 6 Hz was somewhat detrimental.

The exemplar-based engine for estimating posterior probabilities of sub-word units was based on a Lasso solver in a sparse coding procedure. In the Lin18-EMS system we used $17,148$ speech exemplars and $13,504$ noise exemplars. These numbers are about twice as high as the numbers of speech and noise exemplars used in Gemmeke et al. (2011b). The need for large numbers of exemplars in our system is probably related to the combination of features with positive and negative values and the Euclidean distance measure. In Ahmadi et al. (2014) we found that in a sparse coding framework it is advantageous to keep the phase information in the output of the modulation BPFs. The same conclusion was reached by Moritz et al. (2011). However, Baby and Van hamme (2015), who used EMS-like features for training DNNs obtained good results when using only the magnitudes of the amplitude of the output of the modulation filters.

The fact that our EMS features have positive and negative feature values ruled out the use of sparse coding engines based on Kullback-Leibler divergence (the

preferred distance measure in non-negative matrix factorization). It is well known that for many features used in pattern recognition tasks the Euclidean distance does not represent the conceptual distance (e.g., Choi et al., 2014). The default solution is to transform the original features to a space in which Euclidean distance does represent conceptual neighbourhood. We counteracted some of the undesirable effects of the Euclidean distance by the equalization and normalization procedures that we applied to the exemplars and the unknown observations. Forcing all exemplars and unknown observations to unit length makes the Euclidean distance equivalent to cosine distance (Choi et al., 2014). In our equalization procedure the exact same weights are used for the 15 gammatone bands in all $\boldsymbol{M}$ modulation bands. As long as the pattern formed by the magnitude of the 15 numbers in the $\boldsymbol{M}$ modulation bands does not differ substantially between the modulation bands, using fixed weights is beneficial. However, if the patterns become different in some modulation bands because of the different characteristics of the noise, fixed weights can be detrimental. This appeared to be the case with the train station noise in test set B.

We preferred an exemplar-based approach over GMMs or neural networks (including DNNs) for estimating the posterior probabilities, because this approach appears to have closer connections to emerging knowledge about cortical representations of audio signals (Mesgarani et al., 2014a,b) and neural processing. Our research was based on the assumption that some configurations of the modulation filterbank would yield EMS vectors in which a substantial proportion of the features is not affected by the background noise, because the expected values of these features are different for the noise and the speech signals. If the proportion of unaffected features is high enough, the sparse coding engine should be able to match partly damaged EMS vectors to the correct exemplars. That assumption is reinforced by the superior performance of human listeners, especially in tasks where there is little or no help from semantics or world knowledge. The assumption is also in line with widely accepted theories about human pattern recognition, which claim that missing data will be reconstructed (Grossberg and Kazerounian, 2011; Wei et al., 2012; Meyer, 2013). In addition, exemplar-based approaches can handle the very high-dimensional feature vectors produced by the most elaborate versions of the auditory model.

To verify that it is the information in the EMS features, rather than the operation of the sparse coding engine, that drives the performance and to verify that the

findings about the design of EMS features are not limited to a SC procedure for estimating posterior probabilities, we repeated many experiments with the *KNeighborsClassifier* in *scikit-learn* (Pedregosa et al., 2011). We always used the exact same speech-plus-noise dictionaries to 'train' the kNN classifier as were used with the SC engine. We saw the same trend in the results as a function of modulation filterbank configuration in all SNR conditions. For the higher SNR levels the absolute accuracies obtained with the kNN classifier were very close to what we obtained with sparse coding. However, in the lowest SNR levels the SC engine had a clear advantage.

We compared the performance of the Lin18-EMS system with the performance of three other systems on the same data set: Mel-spectra features + SC (Gemmeke et al., 2011b), the MFCC AURORA-2-multi-condition recognizer (Hirsch and Pearce, 2000), and the MFCC ETSI-AFE multi-condition recognizer (Hirsch and Pearce, 2006). In test set A the Lin18-EMS system outperformed the other systems in the lowest SNR conditions. However, the two GMM-based systems outperform the two exemplar-based system by a wide margin in the high SNR conditions. The fact that both exemplar-based system suffered in the same conditions, despite using very different features, shows that the problem is not caused by the EMS features. Also, the lower performance of the exemplar-based systems at the highest SNR levels is not due to the interference of the noise exemplars in the dictionary. In-depth analysis of the activations of the exemplars showed that the noise exemplars receive only very small activations in the highest SNR conditions. Decodings with and without the noise exemplars in the dictionary yielded essentially the same accuracy for clean speech and SNR20 dB. The exemplar-based systems mainly suffer from confusion errors. Moreover, we encountered the same problem with the kNN classifier. It is left to future research to understand what causes the confusions in exemplar-based systems at the highest SNR levels.

It has been shown that the performance of an ASR system can be improved by fusing the posterior probabilities obtained from an exemplar-based system and corresponding estimates from GMM- or ANN-based systems (e.g. Geiger et al., 2013). In Sun et al. (2014) it was shown that fusing the posterior probability estimates of an exemplar-based and a GMM-based system can reduce the word error rate for clean speech in AURORA-2 to less than 0.5%. However, it is unlikely that humans use a similar procedure to accomplish their superior recognition performance.

We also compared the performance of the Lin18-EMS system to the -admittedly few and incomplete- data about human recognition performance on the AURORA-2 task. Using the criteria proposed in Meyer (2013) we found that the performance of our system is more similar to humans than some conventional ASR systems. The only discrepancy is that our system did not show the effect that human accuracy decreases with increasing string length. Our system shares this property with all ASR systems and computational models that do not simulate working memory problems.

In the remainder of this section, we will discuss possible ways to repair some of the weaknesses of the proposed system. First of all, the EMS features might be improved, for example by adding the nonlinear compression that is present in virtually all auditory models, but that was left out in the model of Jørgensen and Dau (2014), because compression was not necessary for the purpose of predicting intelligibility. Including the static 10^{th} power compression in the version of the model in Dau et al. (1996) did increase the recognition accuracy for clean speech, at the cost of a substantial decrease in the SNR-5 dB condition (from 68% correct to 34% correct in test set A). We leave the implementation of the full dynamic compression to future research; we expect that it will show the same positive effect for clean speech without the strong negative effect for the lowest SNR conditions.

The EMS representation of (noisy) speech signals is reminiscent of the approaches advocated in multi-stream ASR architectures (Bourlard et al., 1996a; Tibrewala and Hermansky, 1997; Okawa et al., 1998; Bourlard, 1999; Hermansky and Fousek, 2005b; Hermansky, 2013). A representation in terms of multiple modulation frequency bands is likely to contain features that are not heavily affected by the noise. Instead of designing a procedure to optimally fuse the parallel streams at the feature, the probability or output level, we investigated whether the undistorted features would dominate the distance measure between clean speech exemplars and noisy observations in the sparse coding engine. The recognition accuracy that we obtained on test set A of the AURORA-2 task confirms the viability of this assumption, but the results also show that we are still far from human-like performance in terms of absolute accuracy. The conventional combination of static features, deltas and delta-deltas in ASR corresponds to an auditory model in which the LPF in the modulation filterbank has a cut-off frequency of about 50 Hz. In addition, there is one BPF with a centre frequency of approximately 7 Hz and a quality factor $Q = 1$ and another BPF with a quality factor $Q = 2$.

The fact that conventional ASR systems typically benefit from adding delta-delta coefficients raises the question whether the Lin18-EMS system can be improved by adding $Q = 2$ BPFs with cut-off/centre frequencies at strategically chosen positions. The results of Moritz et al. (2011) provide evidence in support of this assumption.

From recent developments in multi-stream ASR (e.g., Hermansky, 2013) it is clear that it is necessary to combine bottom-up fusion (whether at the level of features, probabilities or outputs) with some kind of knowledge about the best - possibly condition-dependent- way for selecting or combining features. The sparse coding procedure that we used for computing the posterior probabilities of the sub-word units does nothing of the kind. We can see at least two ways in which knowledge could be brought into play. First, it is possible to learn the distributions of individual features or groups of features (per gammatone or per modulation filter) in clean speech from the training material. During test, the likelihood that (groups of) features fit the clean distribution can be estimated, and these estimates can be used as additional weights in computing the Euclidean distances in the Lasso solver. Second, it is possible to improve the conversion of the exemplar activations from the sparse coding procedure to posterior probabilities of sub-word units by involving some kind of learning. In Ahmadi et al. (2014) we argued that we should not aim at the optimal approximation of unknown observations as sparse sums of exemplars; rather, we should aim for the optimal classification of the unknown observations. Research is underway in which we apply label-consistent discriminative dictionary learning to replace the semi-random selection of exemplars by a procedure that learns the exemplars that are optimal for reconstruction and classification (e.g. Jiang et al., 2013).

The need for introducing some kind of learning in the procedure for computing posterior probabilities of sub-word units is strengthened by recent observations of the representation of speech signals in the auditory cortex (Pasley et al., 2012; Mesgarani et al., 2014b). From intra-cranial recordings it can be inferred that representations in auditory cortex still accurately reflect the tonotopic representations formed in the peripheral auditory system. This suggests that speech recognition relies on higher-level processes that operate on the tonotopic representations. These processes can only be successful on the basis of substantial amounts of learning. The need for higher-level operations, including selective attention, is also discussed by Henry et al. (2015), based on the different behaviours of brain

oscillations in frequencies < 5 Hz, which are associated to auditory processing and oscillations at frequencies > 8 Hz, which are associated with higher-level cognitive processing (Luo and Poeppel, 2007).

4.6 Conclusion

In this chapter we investigated to what extent a model of the human auditory system that is capable of predicting speech intelligibility in adverse conditions also provides a promising starting point for designing the frontend of a noise robust ASR system. The long-term goal of the research is to design a computational model that shows human-like recognition behaviour in terms of performance level and the type of errors. We investigated which details of the auditory model configuration are most important for maximizing the recognition performance of an exemplar-based system. We found that a system that combines a frontend based on the envelope modulation spectrum with a sparse coding engine for computing posterior probabilities of sub-word units yields competitive performance as long as the modulation spectrum of the background noise is similar to the noise exemplars in the dictionary. The modulation filterbank must cover the frequency range up to about 20 Hz, but there is no configuration that is optimal for all noise types and all SNR levels. The lower the SNR, the more important becomes a high resolution in modulation frequencies ≤ 6 Hz. Although the accuracy of our system is still below human performance, our system behaves more human-like than MFCC-GMM based ASR systems.

The output of the lowpass filter in the proposed modulation filterbank can be considered as the static features in a conventional ASR frontend, while the bandpass filter outputs can be considered as delta-features which are lowpass filtered with different cut-off frequencies. Using this insight, our results indicate that not only our sparse coding based system, but in fact any classical ASR system, would benefit from a frontend in which the static features, the delta coefficients and the delta-delta coefficients are all represented in a multi-resolution fashion. The highly redundant EMS feature vectors have proven to be a promising starting point for noise robust speech recognition. With a more sophisticated distance measure and

a built-in ability to learn how to use this high dimensional acoustic space to discriminate different sub-word units in different acoustic conditions, an interesting research area opens up where ASR can interface with auditory and brain research.

Chapter 5

Class-Likelihood-Consistent Dictionary Learning for probabilistic classification

Sparse Coding techniques have been used to construct both probabilistic and non-probabilistic classifiers. The accuracy of sparse coders and sparse classifiers can be improved substantially by learning the dictionary from a set of training data. In this chapter we introduce an extension of the K-SVD algorithm for dictionary learning that makes it possible to simultaneously optimize two criteria: the accuracy with which unknown observations can be sparsely coded and the Kullback-Leibler divergence between the output of a probabilistic classifier based on the coder and the class likelihood distributions in the training data. We show the effectiveness of the class-likelihood-consistent K-SVD (CLC-KSVD) learning with three tasks, a sandbox task in which the class likelihood distributions are perfectly known and two speech recognition tasks in which the reference class likelihood distributions must be created. In all tasks CLC-KSVD improves classification criteria at the level of individual observations. In tasks in which the output of the CLC-KSVD classifier is used as input for a back-end decoder, the gain at the application level depends on the capability of the back-end to fully harness the benefits of CLC-KSVD.

5.1 Introduction

In many applications, it is useful to define discrete classes, despite the fact that the signals representing different class members exhibit appreciable overlap in observation space. When classes overlap substantially, there are (at least) two possible approaches to build an adequate classifier: focus on maximizing the proportion of test samples for which the most likely class is the 'correct' one, or build a classifier that yields a maximally correct estimation of the class probabilities. The latter approach is more desirable in applications where the output of a classifier is not the final product, but where the classifier output is used as input for further processing. Typical examples include applications where non-stationary time series must be treated, for example to discover events of interest in lengthy sequences, e.g., Ntalampiras et al. (2011). Automatic speech recognition is another clear example in which retaining only the most likely phone candidate at each point in time is insufficient to achieve an appropriate speech decoding. Instead one should strive for accurate estimates of the posterior probability of a series of runner-up candidates, e.g. Bourlard and Morgan (2012). Classifiers that assign not only the most likely class to some input, but estimate the posterior probability of all possible classes, are often denoted as a probabilistic classifiers. Probabilistic classifiers are widely used in many machine learning application areas, for example in data mining (Friedman et al., 2001; Glickman et al., 2005), computational biology (Horton and Nakai, 1996) and network traffic (Zuev and Moore, 2005). Methods like Naive Bayes classifiers (John and Langley, 1995), logistic regression (Walker and Duncan, 1967), support vector machines (Vapnik, 1995) and multilayer perceptrons (Pal and Mitra, 1992), are the most popular probabilistic classifiers used in machine learning. In many classification tasks, the overlap pattern in the labeled training data can be established. These overlap patterns can then be used to obtain reference class probability vectors that describe the possible between-class confusions in the training data. It would be attractive to use these class probability vectors for supervised training of more advanced, improved classifiers, instead of training those classifiers with unique labels for all training observations. In this chapter we present a method for doing just that. We develop the mathematics for training and testing a probabilistic classifier based on sparse representations in an overcomplete basis (Huang and Aviyente, 2006).

Sparse coding has already been used for sparse classification in order to estimate HMM state posterior probabilities in automatic speech recognition. In Gemmeke et al. (2011b) and Ahmadi et al. (2014), an exemplar dictionary consisting of randomly selected labelled speech tokens is used to decode an observed speech signal. The resulting sparse code is then used to estimate the probability that the observed signal belongs to a specific HMM state. The recognition performance reported in Gemmeke et al. (2011b) and Ahmadi et al. (2014) suggests that the sparse codes obtained from an exemplar dictionary have a reasonable discriminative power, but that there is substantial room for improvement.

Applying a learned dictionary rather than a dictionary with randomly selected atoms, may form a suitable way to increase the discriminative power. Dictionary learning (DL) is a technique to obtain a set of atoms learned from a set of training signals. The aim of learning is to find or construct dictionary atoms that are optimally suited to approximate an observed signal with a linear combination of a small number of atoms. DL-based methods have been proven to be beneficial in many different applications, particularly in the image processing domain (Elad and Aharon, 2006; Mairal et al., 2008b). Conventionally, the DL framework focuses on the best possible signal reconstruction. However, since the learned dictionary provides a sparse representation model for the data space, it also reflects the intrinsic class distribution of data and can therefore be used for classification tasks (Yang et al., 2009; Wright et al., 2009).

Recent research has developed several extensions to DL to adapt it specifically for classification tasks and to improve the classification performance of the learned dictionaries. In these classification-oriented DL methods, either a set of class-specific sub-dictionaries are learned (Yang et al., 2010b; Ramirez et al., 2010) or an overall discriminative dictionary is learned by forcing sparse coefficients to be discriminative (Mairal et al., 2009b; Jiang et al., 2013; Yang et al., 2011). Task-specific discriminative DL, like Jiang et al. (2016) for classifying-machine faults, and the method proposed in Chen et al. (2013) to learn a discriminative dictionary from ambiguously labelled data are other examples of discriminative learning proposed in literature. Almost all above-mentioned works on discriminative learning had their main focus on optimal classification, i.e., classification where the evaluation criterion was only concerned about the correctness of the most likely candidate.

In this chapter we address the question how to learn a dictionary from a set of training data such that the dictionary is optimal for probabilistic classification using sparse coding. For that purpose we need the dictionary to reflect the data structure as well as the class distribution in the data space. Therefore, we will strive for a learning procedure that not only minimizes the reconstruction error, but also aims at accurate estimation of the class membership probabilities for each observation. To ensure this, the DL objective function is constrained to minimize the reconstruction error alongside with some probabilistic classification loss function that measures how well the reference and estimated class probability vectors match. The main contribution of this chapter consists of the way in which this probabilistic loss function is defined and how the resulting learning problem is converted into a practically solvable form. By using an upper bound of the symmetrized Kullback-Leibler divergence (KLD) as a loss function, we show that it is possible to keep using the computationally efficient KSVD framework for dictionary learning when the atoms are not only designed to minimize signal reconstruction error, but also to optimize their performance in a (probabilistic) classification task.

We evaluate the newly proposed classifier in three experiments. In the first one, it is applied to a synthetic data set for which the class overlap in the data space is exactly known. We compare the performance of our algorithm with other probabilistic classifiers in terms of classification accuracy and also the similarity between the underlying and estimated class-class confusion matrix. In the second and third experiment, we apply the proposed probabilistic classifier to two well-known problems in automatic speech recognition.

5.2 Method

Assuming a set of classes $\Theta = \{\theta_1, ..., \theta_L\}$, the goal of probabilistic classification is to estimate the probability that some input vector y belongs to class θ_j ($1 \leq j \leq L$). In Section 5.2.1, we describe how a KSVD learned dictionary can be used for probabilistic classification using sparse codes. Later in Section 5.2.2, we propose our novel class-likelihood-consistent (CLC) dictionary learning and the probabilistic classifier that utilizes the CLC learned dictionary to obtain a more accurate estimation of the class probabilities.

5.2.1 Sparse classification using a KSVD learned dictionary

In this section, we introduce a probabilistic classifier based on the sparse coding approach, in which a dictionary of learned atoms is used. The learning is based on the the KSVD approach introduced by Aharon et al. (2006). Sparse classification can be implemented by means of a KSVD learned dictionary, provided sufficient labelled training data is available from which atom-label correspondences in the dictionary can be inferred. In doing so, we will assume that each feature vector y_i from the training set is associated with a single (golden) class label $\mathcal{L}_{y_i}$. We will first summarize how KSVD dictionary learning takes place when the goal is to accurately reconstruct input feature vectors by means of a sparse, linear combination of atoms. Subsequently, we will show how this learned dictionary can be used to estimate the class probability distribution of some unknown input vector y_i.

5.2.1.1 Reconstructive Dictionary Learning using KSVD

For finding a dictionary $\hat{D}$ that is optimally suited to reconstruct a set of n-dimensional input feature vectors, $Y = [y_1 \dots y_i \dots y_N] \in \mathfrak{R}^{n \times N}$ by means of a linear combination of at most T dictionary atoms, one may use KSVD (Aharon et al., 2006) to solve the following problem:

$$\left\{\hat{D}, \hat{X}\right\} = \arg\min_{D,X} \|Y - DX\|_2^2 \quad s.t.\ \forall i, \|x_i\|_0 \leq T \tag{5.1}$$

where $\|Y - DX\|_2^2$ represents the reconstruction error, $D = [d_1, \dots d_k, \dots d_K] \in \mathfrak{R}^{n \times K}$ is the learned dictionary, $X = [x_1, \dots x_i, \dots x_N] \in \mathfrak{R}^{K \times N}$ are the sparse codes of input signals Y, and T is the sparsity level of X (each x_i has fewer than T nonzero elements). With the number of dictionary atoms $K > n$ the dictionary D will be over-complete, which has the advantage that the reconstruction typically is more robust in the presence of noise, requires fewer dictionary atoms (can be sparser), and is more flexible in matching structure in the data (Lewicki and Sejnowski, 2000). Efficient algorithms are available for solving eq. (5.1). The algorithm proposed in Aharon et al. (2006) solves eq. (5.1) by means of an iterative

method that alternates between sparse coding of the input signals Y based on the current D and updating the dictionary atoms to better fit the data. The KSVD algorithm (with CoefROMP) proposed in Smith and Elad (2013) uses essentially the same approach, but improves the computational efficiency of the two steps. Since in some cases signal exemplars can only have positive values, while the learned dictionary may also contain atoms with negative feature values, in Aharon et al. (2005) a variant of the KSVD algorithm is proposed to handle extraction of dictionaries with non-negative atoms only.

5.2.1.2 Class likelihood estimation using the learned reconstructive dictionary

The learned dictionary $\hat{D}$ and the matrix $\hat{X}$ containing the sparse coefficients of the training data can be used to define an *atom contribution matrix* $\Phi = [\phi_1 \cdots \phi_K]$ with matrix elements $\phi_k(l)$ as:

$$\phi_k(l) = \sum_{i \ s.t. \ \mathcal{L}_{y_i}=l} \hat{X}(k,i)\,, \qquad 1 \le l \le L \tag{5.2}$$

where L is the number of different labels, ϕ_k is the *contribution vector* for the k^{th} atom in the dictionary $\hat{D}$. The L-dimensional vector ϕ_k indicates how strongly the k^{th} dictionary atom is associated with each of the L classes.

Consider a new input observation y of which the class assignment is unknown. Given the dictionary $\hat{D}$, the sparse code $\hat{x}$ that optimally reconstructs y under the sparsity constraints can then be computed using OMP/Lasso (Pati et al., 1993; Efron et al., 2004):

$$\begin{aligned} \hat{x} = \arg\min_x \left\| y - \hat{D}x \right\|_2^2 \quad & s.t. \ \|x\|_0 \le T \ \ (OMP) \\ & (or \ s.t. \ \|x\|_1 \le T \ (Lasso)). \end{aligned} \tag{5.3}$$

Since the vector $\hat{x}$ represents the activation of each dictionary atom in decoding the input observation y, the product of $\hat{x}$ and the atom contribution matrix Φ

can be considered as an estimate of the class likelihood vector of observation y. Consequently, the estimated class likelihood vector $p(y|\Theta)$ can be obtained by a linear combination of all atom contribution vectors using the weights $\hat{x}$ that yielded the smallest reconstruction error. We will use the symbol f to denote this feature-based estimate.

$$\hat{p}(y|\Theta) \approx f = \hat{x}.\Phi \tag{5.4}$$

5.2.2 Class-Likelihood-Consistent Dictionary learning (CLC-KSVD)

Reconstructive Dictionary Learning approaches are designed to produce dictionaries for accurate signal reconstruction. They do not utilize any class information about training signals, hence they are referred to as unsupervised dictionary learning. Although an unsupervised learned dictionary can be used for classification, recent research indicates that dictionaries obtained by supervised learning typically yield better classification performance (Mairal et al., 2008a; Pham and Venkatesh, 2008; Jiang et al., 2013; Yang et al., 2011). Supervised learning can be achieved by adding an extra term to the objective function that not only penalizes reconstruction error, but also incorrect classification. Different approaches have been proposed to introduce such a penalty, e.g., in the form of a softmax discriminative cost function (Boureau et al., 2010; Mairal et al., 2009b, 2008a), Fisher discrimination criteria (Huang and Aviyente, 2006), linear predictive classification error (Pham and Venkatesh, 2008; Zhang and Li, 2010), and a hinge loss function (Yang et al., 2010a; Mairal et al., 2008a). The method used in this chapter is in line with the one in Jiang et al. (2013) and Zhang and Li (2010), in which the objective function of DL is enriched by a label consistency term. However, we enforce class likelihood consistency rather than class label consistency. During learning, our method aims at constructing dictionary atoms in such a way that for each input the reconstruction error is minimized, while at the same time maximizing the similarity of the estimated class likelihood vectors and the 'true' class likelihood vectors of the training observations.

We assume that a 'true' class likelihood vector is available for each input vector y_i in the training set. Let us denote this vector as $q_i = P(y_i|\theta) = [q_i(1), ..., q_i(L)]$,

where $q_i(j) = P(y_i|\theta_j)$ denotes the 'true' likelihood of vector y_i given class $\#j$. As explained in Section 5.2.1.2, the sparse codes x_i can be considered as features for classification of the input vector y_i (see eq. (5.4)). If, in addition, we assume that a classifier $g(x_i, V)$ with parameters (V) is available, which produces likelihood estimates for each class, $p_i = [p_i(1), ..., p_i(L)] = \hat{P}(y_i|\theta)$, we can optimize the classifier parameters such that a classification loss function $\sum_i S(p_i, q_i)$ (summed over all training signals) is minimized. For reasons of mathematical convenience, we will make the additional assumption that g is a linear classifier (Bishop, 2006) that can be described as $g(V.x_i)$. This enables us to propose a joint learning algorithm that optimizes V alongside with D.

Probability distributions are best compared by means of their cross-entropy or the Kullback-Leibler divergence. Therefore, with $P = [p_1...p_N]$ denoting the estimated likelihood vectors corresponding to train signals $Y = [y_1, .., y_N]$ with corresponding 'true' class likelihood vectors $Q = [q_1...q_N]$, the aim of the CLC-KSVD learning procedure is to minimize the symmetrized version of the Kullback-Leibler divergence between the distributions of p_i and q_i pairs, i.e.:

$$S(p_i, q_i) = \sum_{j=1}^{L} (p_i(j) - q_i(j)) \ln\left(\frac{p_i(j)}{q_i(j)}\right) \tag{5.5}$$

Therefore, the CLC dictionary learning problem can be formulated as:

$$\left\{\hat{D}, \hat{V}, \hat{X}\right\} = \underset{D,V,X}{\arg\min} \|Y - DX\|_2^2 + \lambda.S\left(g(V, X), Q\right) \quad s.t.\ \forall i, \|x_i\|_0 \leq T \tag{5.6}$$

where, as before, $\|Y - DX\|_2^2$ represents the reconstruction error, but which is now supplemented with $S\left(g(VX), Q\right)$ representing the probabilistic classification loss. λ is a scalar controlling the relative contribution of the classification loss.

Since KSVD is based on minimizing a Euclidean distance, eq. (5.6) cannot be solved directly using KSVD. In order to be able to keep profiting from the computational efficiency of the KSVD procedure, we need to develop a Euclidean approximation of the KLD. Since $p_i(j)$ and $q_i(j)$ are probabilities, their values are always between zero and one which is the region where the derivative of the ln(.) function is larger than 1. For $0 < p_i(j) \leq q_i(j) \leq 1$, it holds that:

$$\ln p_i(j) - \ln q_i(j) \geq p_i(j) - q_i(j) \geq 0. \tag{5.7}$$

Multiplying the two sides of eq. (5.7) by $(p_i(j) - q_i(j)) > 0$, we get:

$$\xrightarrow{\times(p_i(j)-q_i(j))} (p_i(j) - q_i(j)) \ln\left(\frac{p_i(j)}{q_i(j)}\right) \geq (p_i(j) - q_i(j))^2, \tag{5.8}$$

and by multiplying with $\ln\left(\frac{p_i(j)}{q_i(j)}\right) > 0$:

$$\xrightarrow{\times \ln(\frac{p_i(j)}{q_i(j)})} \left(\ln \frac{p_i(j)}{q_i(j)}\right)^2 \geq (p_i(j) - q_i(j)) \ln(\frac{p_i(j)}{q_i(j)}). \tag{5.9}$$

Thus we may write:

$$\sum_j (p_i(j) - q_i(j))^2 \leq \sum_j (p_i(j) - q_i(j)) \ln\left(\frac{p_i(j)}{q_i(j)}\right) \leq \sum_j (\ln p_i(j) - \ln q_i(j))^2. \tag{5.10}$$

Analogously, it can be proven that eq. (5.10) also holds if $p_i(j) > q_i(j)$. Eq. (5.10) means that the symmetrized KLD between p_i and q_i is always larger than the square of the Euclidean distance between p_i and q_i. The latter could be interpreted as the reconstruction error of the likelihood vectors ($||(p_i - q_i)||_2^2$). In addition, it holds that $||(\ln p_i - \ln q_i)||_2^2$ is an upper-bound for the KLD.

Eq. (5.10) allows us to cast the optimization problem eq. (5.6) into one with the same form as the KSVD problem. By minimizing the upper bound of eq. (5.10) instead of $S(p_i, q_i)$ in eq. (5.5) (and thereby relaxing the constraints), we may redefine the dictionary learning problem eq. (5.6) as:

$$\left\{\hat{D}, \hat{V}, \hat{X}\right\} = \underset{D,V,X}{\arg\min} \|Y - DX\|_2^2 + \lambda \|\ln Q - VX\|_2^2, \quad s.t.\ \forall i, \|x_i\|_0 \leq T \tag{5.11}$$

in which minimizing $\|\ln Q - VX\|_2^2$ minimizes the upper bound of the symmetrized KLD between the p_i and q_i distributions, and where $\ln P = [\ln p_1 ... \ln p_N] = VX$, so that V denotes the probabilistic classifier parameters that can be interpreted as a class likelihood dictionary.

Optimization and classification

The problem in eq. (5.11) can be rewritten as:

$$\{\hat{D}, \hat{V}, \hat{X}\} = \underset{D,W,X}{\arg\min} \left\| \begin{bmatrix} Y \\ \sqrt{\lambda} \cdot \ln Q \end{bmatrix} - \begin{bmatrix} D \\ \sqrt{\lambda} \cdot V \end{bmatrix} X \right\|_2^2 \quad s.t.\ \forall i, \|x_i\|_0 \leq T \tag{5.12}$$

which is now of the form that can be solved using KSVD. In eq. (5.12), which we will call the Class-Likelihood-Consistent KSVD (CLC-KSVD) problem, each column of the matrix $\begin{bmatrix} Y \\ \sqrt{\lambda} \cdot \ln Q \end{bmatrix}$ is a vector $\begin{bmatrix} y_i \\ \sqrt{\lambda} \cdot \ln q_i \end{bmatrix}$ that can be considered as a training observation vector, which consists of a feature part y_i and a class log-likelihood part $(\lambda \cdot \ln q_i)$. In real world applications, such as classification of acoustic filterbank energy features, the feature part y_i often has only positive values and is readily normalized to unit Euclidean norm. This does not hold for the likelihood part, however, since the $\ln q_i$ is an all-negative vector of which the minimum value is defined by the floor value of the likelihood vector. Since we prefer to have a value equal to zero for all elements that do not play a role, both in the feature part (y_i) and in the label part $(\sqrt{\lambda} \cdot \ln q_i)$, we rewrite the objective function of the KLD minimization as:

$$\sum_j (\ln p_i(j) - \ln q_i(j))^2 = \sum_j [(\ln p_i(j) + C) - (\ln q_i(j) + C)]^2 \tag{5.13}$$

Doing so, this allows us to reformulate eq. (5.12) as:

$$\{\hat{D}, \hat{V}, \hat{X}\} = \underset{D,W,X}{\arg\min} \left\| \begin{bmatrix} Y \\ \sqrt{\lambda} \cdot \tilde{Q} \end{bmatrix} - \begin{bmatrix} D \\ \sqrt{\lambda} \cdot V \end{bmatrix} X \right\|_2^2 \quad s.t.\ \forall i, \|x_i\|_0 \leq T \tag{5.14}$$

in which $\tilde{Q} = [(\ln q_1 + C) \ldots (\ln q_N + C)]$ and $V.X$ which now approximates $\tilde{Q}$ will have different values compared to V and X in eq. (5.12). Thus we may interpret $V.X$ as: $V.X = \ln P + C = [(\ln p_1 + C) \ \ldots (\ln p_N + C)]$. By shifting the vector $\ln q_i$ with a constant $C = \ln(min(Q))$, we ensure that the classes with close to zero

likelihood will get a value of *zero* in $\tilde{Q}$. The feature sub-vector Y and the label sub-vector $\sqrt{\lambda} \cdot \tilde{Q}$ are first normalized independently of each other. Then, the complete vector is normalized to length 1.

After the learning, the feature dictionary $\hat{D}$, the class likelihood dictionary $\hat{V}$ and the sparse codes for the training data ($\hat{X}$) are available. Next, substituting $\hat{X}$ in eq. (5.2), the atom contribution matrix Φ can be computed.

With $\{\hat{D}, \hat{V}$ and $\Phi\}$ we effectively have two classifiers. The first classifier is, as before, the one based on the atom contribution matrix Φ: given the sparse code $\hat{x}$ computed from eq. (5.3), an estimate of the class likelihood vector f can be obtained using eq. (5.4). Since the matrix X is optimized using eq. (5.14), the atom contribution matrix Φ of the jointly learned dictionary is expected to be more discriminative. Therefore, we expect to obtain an estimate of f that is more accurate than the one obtained from a dictionary that was only optimized for reconstruction accuracy. The second classifier is based on the likelihood dictionary $\hat{V}$. A log-likelihood vector can be obtained, using $\hat{x}$ computed by eq. (5.3) given the dictionary $\hat{D}$ as

$$\ln q = \hat{V} \cdot \hat{x} - C \quad \Rightarrow \quad q = \exp(\hat{V} \cdot \hat{x} - C) \tag{5.15}$$

Although both classifiers are optimized in the same learning procedure, it is not possible to predict which one provides the most accurate estimate for class likelihoods. Our experiments indicate that the most accurate class likelihood estimates can be obtained by exponential fusion of the estimate f, based on reconstruction of feature vectors and q, based on reconstruction of class likelihood vectors:

$$\hat{p}(y_i|\Theta) = f_i^{\alpha}.q_i^{\beta} \tag{5.16}$$

where α and β can be optimized/tuned on a development set of data.

Having the likelihood of classes given the input signal, the class posterior probabilities can be found using Bayes theorem:

$$\hat{p}(\Theta|y_i) \approx \hat{p}(y_i|\Theta).p(\Theta) \tag{5.17}$$

where $p(\Theta)$ is the vector of class priors.

In applications where the prior probability of classes are (almost) equal, a uniform sampling of the data space is appropriate for creating the set of training signals. If there is a significant difference in the number of members per class, uniform sampling may lead to under-representation of small classes. In such a case, it is preferable to create a balanced training data set in which an equal number of signals from each class is included. This will allow the learning of a more accurate classifier (Bishop, 2006). Although using a balanced training set is tantamount to modifying the observation space, the effect of the distortion can later easily be compensated using eq. (5.17).

5.3 Experiments

We evaluate our proposed classification approach in four experiments. We start the evaluation with a task in which the data set is generated using a predefined model. Therefore, oracle knowledge about the distribution and overlap of the data classes is available. After analysing the performance of the proposed method on data with a simple distribution and overlap pattern, we conduct two experiments on clean speech recognition. The first one is on a small speech recognition task in which we estimate sub-word posterior probabilities. In the second one, our probabilistic classifier is applied to a phone classification task. We conduct a fourth experiment on noisy speech recognition using artificially noisified speech utterances. In the experiments with speech data there is no true distribution of the classes in the observation space available to define the true likelihood. Therefore, the reference likelihoods of the training and test observations must be inferred from knowledge about the task and/or training data. The main focus in the experiments is on the improvement that is obtained by CLC-KSVD, relative to the original KSVD-based classifier. To be able to estimate the relevance of the improvement, we will also compare the results with estimates obtained from a sparse classifier that uses an exemplar dictionary and a Naive Bayes classifier (John and Langley, 1995). To implement the Naive Bayes classifier, we used the `naive Bayes` object in Matlab, which does not require parameter tuning on a development set. For the noisy speech recognition experiment we include the results with estimates obtained from a multilayer perceptrons (MLP) classifier as a comparison reference system.

In the evaluations we will use several related -but different- criteria and measures. The classification accuracy of each classifier is reflected in %correct. The average KLD between the estimated and the reference class probabilities is used to measure the probabilistic classification performance. To have an indication of how the probability mass is divided between correct and wrong classes, we use the C-index, which is defined as the average ratio between the probability mass assigned to the correct class and the total probability mass assigned to all other classes. For example, a C-index=80 means that in the estimated class probability vectors, on average 80% of the probability mass is assigned to correct classes. By comparing estimated and true class assignments for a set of test data points, we can form the confusion matrix. If we have oracle knowledge about the data distribution, we can compare the estimated confusion matrix with the confusions computed from the known distribution. If the model that generated the data is known, it is possible to compute a measure of the physical overlap between the classes (cf. section 5.3.1). From that measure we can obtain an estimate of the upper bound of the classification accuracy in terms of %correct that a classifier can obtain.

We form the initial dictionary D_0 for (CLC-)KSVD learning by randomly selecting vectors from the data space. The initial dictionary will then be an exemplar dictionary, which can be used for sparse classification as in Gemmeke et al. (2011b); Ahmadi et al. (2014). Since the initial dictionary D_0 is a set of labelled data points, the atoms d_i have a unique label $\mathcal{L}_{d_i}$. We can define the atom contribution matrix of the initial dictionary as:

$$\Phi_0(i,j) = \begin{cases} 1 & \mathcal{L}_{d_i} = j \\ 0 & otherwise \end{cases} . \tag{5.18}$$

To classify a test data point y using the initial dictionary, we first compute its corresponding sparse code $\hat{x}_0$ by substituting D_0 in eq. (5.3). The likelihood vector is then computed by substitution of Φ_0 and $\hat{x}_0$ in eq. (5.4).

In all three experiments, a train data set is used to learn a KSVD-based reconstructive dictionary by solving eq. (5.1); we refer to this dictionary as $\hat{D}$. We compute the corresponding atom contribution matrix Φ and use eq. (5.3) and eq. (5.4) to estimate the class likelihood vector of the test data points. We also use the train data together with their class probability vectors to jointly learn

the CLC-KSVD dictionaries $\hat{D}_{clc}$ and $\hat{V}$ via eq. (5.14). We then use eq. (5.2) to compute the atom contribution matrix Φ_{clc}. The sparse codes for the test data are obtained by substituting $\hat{D}_{clc}$ in eq. (5.3). The test data is then classified both using eq. (5.4) (the f-estimate) and (5.15) (the q-estimate). We use a development set to tune the regularization factor λ in eq. (5.14) to maximize the classification accuracy of the f-estimate and of the q-estimate. The class likelihood estimate of the test data points is then obtained using the fusion in eq. (5.16). To tune α and β in eq. (5.16), a grid search using the development set is conducted.

5.3.1 Evaluation on synthetic data

The data set on which we start the evaluation consists of two-dimensional data points, randomly generated by a three component Gaussian mixture model:

$$p(y) = \sum_{i=1}^{3} p(\theta_i)\mathcal{N}(\mu_i, \Sigma_i) \tag{5.19}$$

where:

$$\begin{cases} p(\theta_i) = 1/3 \ ; \ i = 1, 2, 3 \\ \mu_1 = [7, 15] \ ; \ \Sigma_1 = \begin{bmatrix} 5 & 0 \\ 0 & 0.5 \end{bmatrix} \\ \mu_2 = [9, 13] \ ; \ \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \\ \mu_3 = [12, 1] \ ; \ \Sigma_3 = \begin{bmatrix} 6 & -1.5 \\ -1.5 & 1 \end{bmatrix} \end{cases} \tag{5.20}$$

The three distributions defined by eq. (5.20) form two overlapping classes, and one separate class (see Figure 5.1). The data set is designed this way for two reasons. First, we want to see how well CLC-KSVD is able to distinguish the solitary class from the overlapping classes. Second, we want to see to what extent the confusions between data points from the overlapping classes made by the classifiers reflect the uncertainty that is inherent in the generative model. We generated a train and a development set, each containing 1500 data points, by sampling from

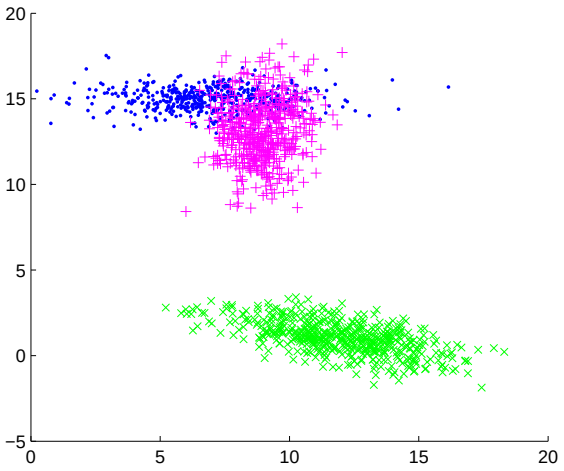


FIGURE 5.1: Distribution of the 3-class synthetic data. blue: $\{\mu_1, \Sigma_1\}$; magenta: $\{\mu_2, \Sigma_2\}$, green: $\{\mu_3, \Sigma_3\}$.

the three individual distributions. By doing so, we know the true class membership of all samples. A test set containing 3000 data points is generated in the same way. Initial over-complete dictionaries with 7 atoms from each of the three classes are generated in the same way. Using eq. (5.19), the true likelihood of each data point y_i, given class $\#j$, is computed as:

$$q_i : q_i(j) = p(y_i|\theta_j) = \frac{1}{2\pi\sqrt{|\Sigma_j|}} \exp\left(-(y_i - \mu_j)^T \Sigma_j^{-1} (y_i - \mu_j)\right) \tag{5.21}$$

The train data set is then used to train the reconstructive dictionary $\boldsymbol{D}$ using KSVD. A CLC-KSVD dictionary is also learned using the training data together with their true likelihood vectors. Using the development set, we searched for the optimal value of the regularization factor λ in the range$\{0.0001, \cdots, 1\}$ with exponentially increasing step size. With the optimal value of $\lambda = 0.8$ we obtained optimal fusion weights $\alpha = 5$ and $\beta = 8.5$. These values minimize KLD, while the classification accuracy is close to maximum. The global optimum for KLD and %correct do not necessarily coincide, because it is possible that the highest posterior assigned to data points located in the overlap region is for the 'wrong' class, while the class probability vector is still close to the true one. It is also possible that points are assigned the correct class label, despite a large distance between the estimated and the true probability vectors.

	% correct	C-index	KLD	Confusion matrix (CM)
Exemplar (Initial Dic.)	83.43	81.32	7.91	$\begin{bmatrix}0.87 & 0.39 & 0.04\\0.10 & 0.61 & 0.00\\0.02 & 0.00 & 0.96\end{bmatrix}$
KSVD	87.80	80.60	1.08	$\begin{bmatrix}0.73 & 0.27 & 0.00\\0.27 & 0.71 & 0.02\\0.00 & 0.02 & 0.98\end{bmatrix}$
CLC-KSVD	88.20	87.89	0.67	$\begin{bmatrix}0.78 & 0.14 & 0.00\\0.21 & 0.85 & 0.00\\0.00 & 0.00 & 0.99\end{bmatrix}$
Naive Bayes	88.07	86.15	0.26	$\begin{bmatrix}0.77 & 0.18 & 0.00\\0.23 & 0.82 & 0.00\\0.00 & 0.00 & 0.99\end{bmatrix}$
Reference	UB=90.79	—-	0	$\begin{bmatrix}0.84 & 0.09 & 0\\0.16 & 0.91 & 0\\0 & 0 & 1.00\end{bmatrix}$

TABLE 5.1: Classification performance on synthetic data. Reference values are computed based on the generative model

To be able to assess the differences between the classifiers in terms of %correct, we use the oracle knowledge about the data space to compute an upper-bound for the classification accuracy, which depends on the intrinsic overlap between classes. Let y_i be a point with class label $\mathcal{L}_{y_i}$. There are three classes $\{\theta_1,\theta_2,\theta_3\}$, so $\mathcal{L}_{y_i}$ is member of {1, 2, 3}. From Bayes it follows that:

$$P(\theta_k|y_i) = \frac{P(y_i|\theta_k).P(\theta_k)}{\sum_j \big(P(y_i|\theta_j).P(\theta_j)\big)} \tag{5.22}$$

The sum

$$\begin{aligned} UB = \frac{1}{3}\Big(& \sum_{y_i \ s.t. \ \mathcal{L}_{y_i}=1} P(\arg\max_k P(\theta_k|y_i) = 1) \\ & + \sum_{y_i \ s.t. \ \mathcal{L}_{y_i}=2} P(\arg\max_k P(\theta_k|y_i) = 2) \\ & + \sum_{y_i \ s.t. \ \mathcal{L}_{y_i}=3} P(\arg\max_k P(\theta_k|y_i) = 3)\Big) \end{aligned} \tag{5.23}$$

is the upper bound for the expected classification accuracy, and $1 - \boldsymbol{UB}$ is the physical class overlap within the data. We can also compute a reference confusion matrix CM in which the CM(j,k) reflects the probability of assigning a class j label to a token that is labeled as class k, i.e. $p(\theta_j|\theta_k)$ for $k \neq j$. And CM(j,j) is $p(\theta_j|\theta_j) = 1 - \sum_{k\neq j} p(\theta_j|\theta_k)$.

Table 5.1 shows the results of four classifiers, viz. based on the exemplar dictionary, the dictionaries learned with KSVD and CLC-KSVD, as well as a Naive Bayes classifier. The columns show %correct, the C-index, the KLD and the confusion matrices. From the Table it can be seen that learning has a substantial effect on classification accuracy and also reduces the KLD between the estimated and the true likelihood vectors. Despite the higher %correct, the C-index obtained with the KSVD classifier is slightly lower than with the exemplar dictionary. That is due to the fact that the vectors in the atom contribution matrix of the exemplar dictionary (defined in eq. (5.18)) have a smaller average entropy than the vectors in the atom contribution matrix of the KSVD dictionary (defined in eq. (5.2)). The CLC-KSVD classifier outperforms the KSVD classifier both in terms of KLD and %correct, indicating that the class likelihood consistency constraint not only improves the likelihood estimation, but it also increases the classification accuracy. In comparison to the Naive Bayes classifier, the CLC-KSVD classifier performs slightly better in terms of accuracy, but not KLD.

The most direct information for answering the questions whether CLC-based classifiers show superior performance in separating the isolated class and in accounting for the overlap between the two other classes is provided by the confusion matrices. The Naive Bayes and CLC-KSVD classifiers are trained in a supervised procedure, unlike the other two classifiers. As a result, they are more successful in separating the solitary class from the two other classes. The confusion matrix of CLC-KSVD is closer to the matrix in the reference row than the Naive Bayes confusion matrix. This is mainly because CLC-KSVD is more accurate in estimating overlap between overlapping classes. Specifically, for CLC-KSVD, the entries $\hat{p}(\theta_1|\theta_2)$ and $\hat{p}(\theta_2|\theta_1)$ in the confusion matrix are closer to $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$ in the reference confusion matrix. The comparisons between the confusion matrices can also be quantified using the RMS error. The RMS error between confusion matrix of the CLC-KSVD classifier and the reference confusion matrix is 0.037, which compares favourably with 0.055 of Naive Bayes; for the other two classifiers the RMS error is larger than 0.1.

From the results of this sandbox example we can conclude that the CLC-KSVD classifier does indeed increase the accuracy in estimating class probabilities, both in terms of handling overlapping and separating non-overlapping classes. It results in close-to-ceiling classification accuracy. At the same time, it reduces the KLD between the estimated and true probability vectors, which we believe is an important property in real-world probabilistic classification tasks. It is interesting to see whether the advantage of CLC-KSVD generalizes to real-world tasks.

5.3.2 Evaluation on clean speech recognition tasks

Automatic speech recognition (ASR) is a textbook example of an application in which classifiers are used for estimating the posterior probability of sub-word units (usually states in a hidden Markov model), rather than for identifying a unique winner. Modern large vocabulary ASR systems use several thousand sub-word units. The observations live in a high-dimensional space, and there is substantial overlap between sub-word units. To investigate whether the advantage of CLC-KSVD generalizes to applications that are much more complex than the toy example in section 5.3.1, we conduct experiments on a word recognition and also a phone recognition task.

5.3.2.1 Small vocabulary word recognition on AURORA-2

In this experiment, we evaluate the CLC-KSVD classifier on AURORA-2 (Hirsch and Pearce, 2000), a small vocabulary task in which the utterances are sequences of connected digits that comprise between one and seven digits. In the conventional approach to the AURORA-2 task, each of the eleven digit words ("oh", "zero", $\cdots$, "nine") is modelled as a sequence of 16 states. In addition, there are three silence states, making for a total of 179 states (or classes). From the fact that even the word "oh" is modelled as a sequence of 16 states it is clear that not all states are acoustically distinct; on the contrary, there is considerable overlap between neighbouring states. Moreover, the word-initial and word-final states are hardly distinguishable from silence states.

We use this knowledge about the structure of the state sequences to define our reference (true) likelihood vectors for train and evaluation data. In the reference

Classifier		Classification accuracy				% Word recognition accuracy
		% Correct	%Correct@3	C-index	KLD	
Exemplar (Initial Dic.)		32.26	52.36	19.20	34.43	94.14 (±0.40)
KSVD		30.37	55.63	13.91	25.98	95.06 (±0.37)
CLC-KSVD	α, β					
	0.4, 0.1	32.00	60.14	14.25	22.92	96.16 (±0.33)
	0.9, 0.3	31.16	56.68	14.79	18.24	95.82 (±0.33)
Naive Bayes		15.60	31.91	15.71	50.16	68.31 (±0.81)

TABLE 5.2: Frame classification and word recognition accuracy on AURORA-2 clean test set. The recognition accuracy data include the 95% confidence intervals.

likelihood vector of frames from word-internal states, the golden label of each frame gets the highest likelihood and the neighbouring states and second neighbours (if existing) get a likelihood of 50% and 25% of the likelihood of the golden label, respectively. In the reference likelihood vectors of frames from word-initial and word-final states the probability mass is uniformly distributed among all bordering states and the 3 silence states. The same holds for the reference likelihood vectors of silence frames. Thus, the reference vectors were defined solely on the basis of prior knowledge of the task, independent of any actual training data.

The prior probability of the 3 silence states depends on the operation of a voice activity detector or the manual segmentation of the utterances. The prior probabilities of the 176 speech states differ only marginally. For this reason we did not include the priors in the AURORA-2 experiment (cf. eq. (5.17)).

We use the envelope modulation spectrum (EMS) feature extraction procedure proposed in Chapter 4 to convert the speech signals into sequence of 285-dimensional feature vectors. We obtain an EMS feature frame for every 10 ms of speech signal. In Chapter 4, an exemplar-based approach was used to estimate the posterior probabilities of the speech states. Sparse codes of input EMS feature vectors are computed, based on an exemplar dictionary consisting of almost 17,000 speech frames, randomly selected from the training data. The procedure for computing posterior probability vectors was similar to the procedure defined by equations (5.18), (5.3) and (5.4), which we used in classification using the initial dictionaries. In Chapter 4, we reported the recognition accuracies obtained by

running a Viterbi search over the likelihood estimates both for clean and noisified speech. In this experiment, we take the results on clean speech signals (cf. Table 4.1) as the baseline and investigate how much we can improve the recognition performance on clean speech signals by replacing the exemplar dictionary with dictionaries learned using KSVD and CLC-KSVD (of the same size) and the corresponding likelihood estimation procedures (obtained from eq. (5.4) and (5.16) respectively).

We train our KSVD and CLC-KSVD dictionaries using a set of $80,000$ EMS feature frames. The training frames are semi-randomly selected from the 8440 utterances in the clean train data set, so that there are almost equal numbers of frames taken from each of 179 classes. An evaluation set of $80,000$ frames is selected in the same way as the training set and used to tune λ, α and β. We obtained $\lambda = 0.002$ as an optimum value for this task. The learned dictionaries are then used to estimate the posterior probability of the HMM-states for the AURORA-2 clean test set. A Viterbi decoder is used to find the best path in the resulting lattice spanned by the 179 states and we then calculate the average word recognition accuracy. We also compute the percentage of correctly labelled frames as an analogue to classification accuracy. In case of classification errors, the correct class might still have a high probability and make for an attractive candidate in the Viterbi search. Therefore, we also compute *%correct@3*, i.e., the percentage of frames for which the correct class is among the three classes with highest probabilities. The C-index and average KLD between the estimated and true likelihood vectors are also computed. The results are summarized in Table 5.2.

The %correct measure obtained with the KSVD dictionary is lower than the %correct measure with the initial dictionary (Exemplar). However, KSVD shows a substantial advantage in %correct@3. This indicates that the probability that the correct class is among the most probable classes is higher with the KSVD dictionary. Moreover, the average KL-distance between the estimated and true likelihood vectors is much lower for the KSVD dictionary. As a result, the word recognition accuracy has increased. This result confirms our expectation that for a task like speech recognition, more accurate estimates of class (state) likelihoods are more important than the classification accuracy of individual frames.

For the CLC-KSVD dictionary, there are two sets of results corresponding to different tuning of the fusion parameters α and β that yielded the highest %correct

or the lowest KLD. In the AURORA-2 task, both local optima yield almost the same word recognition accuracy. This suggests that KLD is at least as powerful a predictor of the performance of the back-end as %correct. The fact that the recognition accuracy with the CLC-KSVD dictionaries is significantly higher than with the KSVD dictionary shows that label consistency is relevant for the back-end.

While the Naive Bayes classifier had competing performance in terms of %correct and KLD on the sandbox task, it is clearly inadequate in the word recognition task. This is due to the fact that the underlying distributions in the AURORA-2 task are not Gaussian. In addition, with 285-D features the Naive Bayes classifier suffers from the curse of dimensionality.

From the results pertaining to the C-index, it can be inferred that the estimated likelihood vectors obtained with the exemplar dictionary are crisper than the vectors obtained with the learned dictionaries. CLC-KSVD yields a slightly higher C-index than KSVD, suggesting that the dictionary learned with the additional constraint results in slightly crisper likelihood vectors. However, the C-index only reflects the posterior vectors of the frames for which the most probable label was the correct one. Therefore, the C-index cannot be expected to predict word recognition accuracy.

5.3.2.2 TIMIT phone classification

In the third experiment, we use CLC-KSVD in the phone recognition task in the TIMIT acoustic-phonetic corpus (Lamel et al., 1989). TIMIT was constructed to train and evaluate speaker-independent phone recognizers. It comprises 630 speakers, who each read aloud 10 sentences. In the original transcriptions 64 phone labels were used. Most experiments collapse this to 48 labels (Lee and Hon, 1989) or 39 (Halberstadt and Glass, 1997) labels, by merging phones that are very similar. In this experiment we use the 48 label list of Lee and Hon (1989) as input for the back-end, while the accuracy of the back-end is computed based on the 39 label list.

The observation vectors are 26-D Mel spectrum features computed using HTK (Young et al., 2009), enriched by the first and second derivative features (Δ and

$\Delta - \Delta$). We use the train set of the TIMIT corpus to create the exemplar dictionary; we also construct sets of training and development frames from the training utterances. The selection of feature frames for the exemplar dictionary, the training and the development set is done by a semi-random procedure, such that there are almost equal numbers of frames from each of the 48 classes. We also made sure there is no speaker overlap between the train and development set. The training and development sets consists of $100,000$ feature vectors. The exemplar dictionary contains $32,000$ speech frames.

Given the time-aligned transcriptions of the data set, the golden class label for each feature frame is available. But there is no golden class likelihood reference provided by the data set. Unlike the AURORA-2 task, it is not possible to infer likelihood reference vectors from some pattern in the sequences of phones in the sentences. Almost all phones can follow all other phones. We computed the overlap between the phones in the training data in the 78-D acoustic space. All vowels can be confused with all other vowels, and the same holds for the consonants. In addition, there are confusions between some vowels and some consonants. Reference likelihood vectors that mirror the confusion patterns appeared to be too fuzzy to have an appreciable effect in CLC-KSVD learning. Therefore, we created reference likelihood vectors by transforming the acoustic distance patterns to make them more crisp, in the sense that the likelihood of the 'golden' label was increased relative to all other labels. We fitted a single Gaussian distribution to all frames in the TIMIT training data that belong to one of the 48 classes, and computed the matrix $G = [g_{jk}]$ of the Euclidean distances between the cluster centres. The columns of G were normalized by dividing all elements in a column vector by the largest element in that vector. Finally, the class likelihood vector of a frame y_i is computed as

$$q_i : q_i(j) = p(y_i|\theta_j) = \begin{cases} 0.75 & \mathcal{L}_{y_i} = j \\ 0.75/g_{jk} & \mathcal{L}_{y_i} = k \end{cases}. \tag{5.24}$$

q_i is then normalized to sum to 1.

Using the reference vectors defined by eq. (5.24) we trained CLC-KSVD dictionaries and used eq. (5.4) and (5.16) to estimate the class (phone) likelihoods for the speech frames of the TIMIT test set. To provide a baseline, we also used the initial (exemplar) dictionary for probabilistic classification of the test data. We

Classifier		Classification accuracy				% phone recognition accuracy
		% Correct	%Correct@3	C-index	KLD	
Exemplar (Initial Dic.)		36.25	59.44	24.20	26.48	54.8 (±1.12)
KSVD		37.64	63.90	13.47	21.87	52.4 (±1.14)
CLC-KSVD	α, β					
	5.9, 0.1	37.92	64.20	32.62	17.99	49.2 (±1.63)
	0.9, 0.3	37.76	63.69	17.01	12.88	52.3 (±1.14)
Naive Bayes		25.40	44.87	19.06	28.71	29.5 (±1.04)

TABLE 5.3: Frame classification and phone recognition accuracy on TIMIT. The recognition accuracies include the 95% confidence intervals.

also trained a KSVD and a Naive Bayes classifier. The regularization factor λ is set to 0.0008 during a tuning on the evaluation set. The frame-based classification results on the test set of TIMIT are shown in Table 5.3.

The first observation from the table is that the frame-based classification accuracy improves thanks to learning. The performance of the CLC-KSVD classifier is shown for two different $\{\alpha, \beta\}$ pairs. One pair yields the maximum %Correct and the other the minimum KLD. While the difference in %Correct between the two $\{\alpha, \beta\}$ pairs is marginal, the pair with the values $\{\alpha = 0.9, \beta = 0.3\}$ yields a substantially smaller KLD. From the C-index data it can be seen that the CLC-KSVD classifier with the $\{\alpha = 5.9, \beta = 0.1\}$ pair that yields the maximum %Correct appears to increase the average proportion of the probability mass attributed to the correct class. The C-index is even higher than for the exemplar-based classifier. There are no substantial differences between the (CLC-)KSVD classifiers in terms of %Correct@3; all three improve this measure relative to the exemplar-based classifier.

We need a back-end to process the probabilistic frame-based classification output to obtain the final phone recognition. Unlike AURORA-2 there is no commonly used back-end. We adapted the TIMIT script in the kaldi package (Povey et al., 2011) such that we could compose the posterior probability matrices generated by our classifiers with the TIMIT finite state transducer that is generated for the conventional situation in which the acoustic information is derived from likelihoods obtained from a GMM. In the conventional approach phones are modelled by a sequence of three states, representing the transitions into and out of the phone and

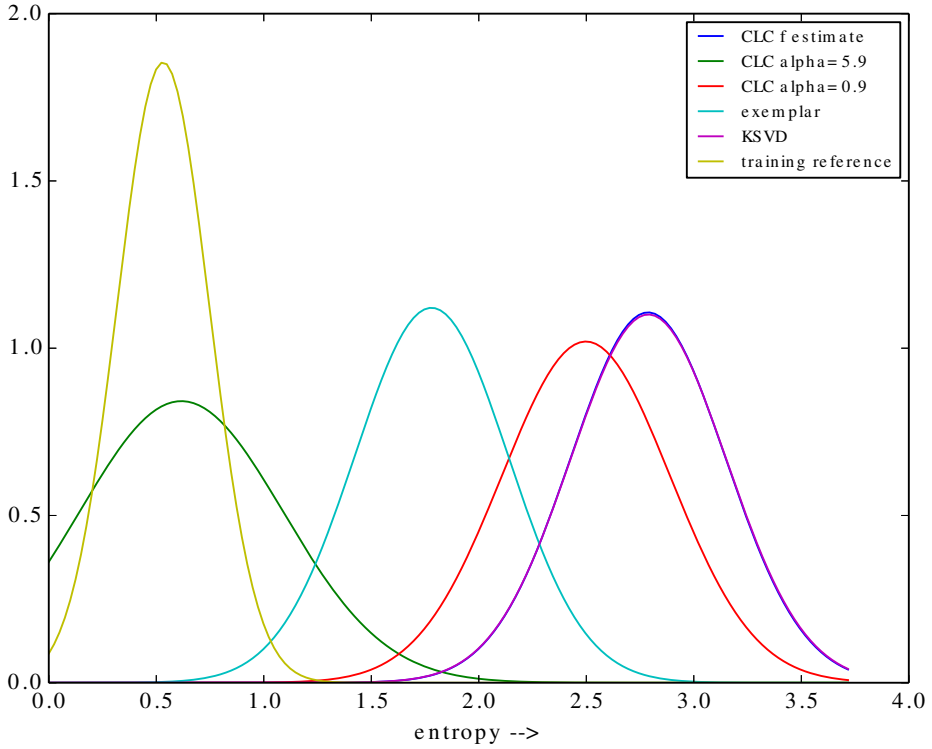


FIGURE 5.2: Distribution of the entropy in the posterior vectors generated by five classifiers, plus the distribution of the entropy in the reference vectors. The curve labelled 'CLC f estimate' refers to the f estimates in eq. (5.4).

the middle part of the phone. In our decoder we used only one state per phone. The back-end accounts for the prior probability of the phones.

From the phone recognition accuracies in the rightmost column of Table 5.3 it appears, unsurprisingly, that frame-based classification accuracy is not a good predictor of the performance of the back-end decoder. The fact that KLD is not a good predictor either, is more difficult to interpret. There are at least two issues, which are difficult to disentangle. The definition of the reference likelihood vectors in on eq. (5.24) that determines the KLD is wrong, or KLD cannot predict the performance of the back-end.

One feature that is somewhat indicative of the effectiveness of class likelihood vectors in speech recognition back-ends is the distribution of their entropy (Sun et al., 2014). Figure 5.2 shows the distributions of the entropy of the posterior estimates of the classifiers in Table 5.3. The entropy distributions of the KSVD classifier and the CLC f-estimate overlap almost perfectly. This is in line with

the observation that the atom contribution matrix Φ resulting from optimizing eq. (5.11) differed only marginally from the corresponding matrix obtained with KSVD learning. The fusion of the atom contribution matrix and the likelihood dictionary in eq. (5.16) results in classifiers whose outputs have lower overall entropy, and a larger standard deviation. However, the fairly low average entropy in the fusion with $\{\alpha = 5.9, \beta = 0.1\}$ is mainly due to the large exponent on α. Yet, it is evident that learning with CLC shifts the entropy distributions in the direction of the distributions imposed by the reference likelihoods vectors. The low average entropy of the reference likelihoods is due to the constant 0.75 in eq. (5.24).

The entropy in the output of KSVD training and the f-estimate in CLC-KSVD is substantially higher than the entropy in the output of the SC classifier that uses the exemplar dictionary. This was to be expected. The atoms in the exemplar dictionary represent a single class. Optimizing the dictionary for maximizing reconstruction accuracy inevitably leads to a dictionary in which atoms are associated with multiple classes. However, entropy alone cannot explain phone recognition accuracy. We ran a phone decoding with an $\{\alpha, \beta\}$ pair that yields an entropy distribution that is very similar to the distribution of the exemplar-based classifier, but that did not yield a better phone recognition accuracy.

5.3.3 Evaluation on noisy speech recognition

In Section 5.3.2 we showed that our proposed probabilistic classifier is effective in estimating speech unit posterior probabilities and that the use of a CLC-learned dictionary leads to an improved recognition performance compared to both an exemplar dictionary and an unsupervised KSVD learned dictionary. In this section we examine how the CLC-learned approach performs in estimating posterior probabilities in noisy speech. For that purpose we use the noisy speech provided in the AURORA-2 data set. The data set contains a multi-condition train set in which the clean training speech signals are artificially noisified by mixing them with four noise types; the noise levels are varied in steps of 5 dB resulting in SNR levels of 20 down to 5dB. In addition there are two test sets with SNR levels of 20 down to −5dB. In test set A the same noise types are present as in the multi-condition training set. Test set B is designed for evaluation of mismatched noise conditions and contains the same utterances as test set A, but these are noisified with four noise types that are not present in the training set. As in Section 5.3.2.1, we

use 285-dimensional EMS features. We take the recognition performance obtained with speech+noise exemplar dictionaries reported in Chapter 4 (c.f. Table 4.1) as the baseline results.

To decode the noisy speech, we consider two approaches. In the first approach, we repeat the exact same procedure as in Section 5.3.2.1, i.e., the posterior probabilities are computed using our probabilistic classifier employing KSVD or CLC-KSVD learned dictionaries. In the second approach, we use a combined speech+noise dictionary constructed by concatenating a learned dictionary (KSVD or CLC-KSVD) and the exemplar noise dictionary used in Chapter 4. The sparse codes of the test speech frames are computed by replacing the combined dictionary in eq. (5.3). After discarding the weights assigned to the noise exemplars, the remaining weights in the sparse codes are used to estimate the state posterior probabilities. We run these experiments for the KSVD and CLC-KSVD learned dictionaries used in the experiment of Section 5.3.2.1. Furthermore, to examine the effect of the training material on noise robustness, we also train a KSVD and a CLC-KSVD dictionary using the AURORA-2 multi-condition instead of the clean training data. The dictionary size, the value of λ and the number of training observations were the same as the clean trained dictionary. These multi-condition trained dictionaries were then used to compute a set of alternative posterior probability estimates. Finally, doing a Viterbi search on each set of estimated posteriors, the word recognition accuracy for each noise-SNR subset of the test data was computed. The values of α and β were tuned during a grid search for each noise-SNR subset. Depending on noise type and SNR level, we obtained optimal values of $0.4 < \alpha < 4$ and $0.1 < \beta < 0.6$. However, we observed that the search landscape around the optimum was always fairly flat. The recognition accuracies averaged over all four noise types for test set A and B are reported in Tables 5.4 and 5.5 respectively.

Let us now compare Tables 5.4 and 5.5 to evaluate the effect on recognition performance of learning algorithm (KSVD or CLC-KSVD), including or not including noise exemplars in the dictionary, and of the type of training data (clean or multi-condition).

KSVD VS. exemplar dictionaries: On test set A, the recognition accuracy obtained using the KSVD learned dictionary is inferior at all SNR levels, no matter if noise exemplars are included or not. On test set B however, the best performing

Classifier	Train	+ Noise exemplars?	SNR (dB)						Avg.
			20	15	10	5	0	-5	
Exemplar		Yes	**94.12**	**93.74**	**91.70**	87.24	79.56	60.51	84.48
KSVD	Clean	No	92.70	91.76	89.10	82.05	62.30	30.78	74.78
		Yes	92.01	91.07	89.39	85.49	76.85	59.15	82.33
	Multi	No	92.69	91.91	90.35	86.23	74.27	45.70	80.19
		Yes	92.01	91.38	90.05	86.36	78.13	60.22	83.02
CLC-KSVD	Clean	No	**94.44**	93.05	90.78	83.51	65.84	34.69	77.05
		Yes	93.90	93.12	**91.84**	**88.18**	**80.50**	**63.99**	**85.25**
	Multi	No	93.58	92.93	**91.40**	87.34	74.98	46.95	81.20
		Yes	93.08	92.56	**91.50**	**88.14**	**80.51**	**63.54**	84.89

TABLE 5.4: Word recognition accuracy on AURORA-2 noisy test set A. In the third column "Yes" refers to speech+noise dictionary and "No" refers to speech dictionary. Bold numbers indicate the best performance per column (within the 5% confidence interval)

Classifier	Train	+ Noise exemplars?	SNR (dB)						Avg.
			20	15	10	5	0	-5	
Exemplar		Yes	**93.10**	**92.50**	89.40	79.78	58.46	27.12	73.39
KSVD	Clean	No	92.06	90.75	88.06	79.13	58.66	29.91	73.09
		Yes	90.72	89.18	85.89	78.94	60.62	33.87	73.20
	Multi	No	91.75	91.20	89.14	83.27	67.51	38.22	76.85
		Yes	90.62	89.47	87.28	80.76	64.29	35.74	74.69
CLC-KSVD	Clean	No	**93.45**	**92.54**	89.09	80.94	60.94	31.72	74.78
		Yes	92.39	91.47	88.92	82.85	67.22	38.44	76.88
	Multi	No	92.71	**92.22**	**89.99**	**84.34**	**68.09**	38.60	**77.66**
		Yes	91.62	90.92	88.66	83.61	**68.59**	**39.78**	77.19

TABLE 5.5: Word recognition accuracy on AURORA-2 noisy test set B. In the third column "Yes" refers to speech+noise dictionary and "No" refers to speech dictionary. Bold numbers indicate the best performance per column (within the 5% confidence interval)

KSVD outperforms the exemplar approach at SNR levels below 10dB and also in terms of average performance over all SNRs. This indicates that, compared to the exemplar dictionary, the KSVD learned dictionary generalizes better to unseen noise types and therefore it is less sensitive to noise mismatch conditions. From Table 5.5 it is clear that the superior performance of KSVD over the exemplar approach at the lowest two SNR levels under the mismatched condition does not depend on the type of training data nor on whether or not noise exemplars are added to the dictionary. Tables 5.4 and 5.5 also show that the best performance of the KSVD approach on both test set A and B is obtained when the dictionary is

trained with multi-condition data and noise exemplars are involved in the sparse coding.

CLC-KSVD VS. KSVD: CLC-KSVD outperforms the KSVD under all test conditions and at all SNR levels. This indicates that the class likelihood consistency constraint is helpful under all conditions, although under some conditions the advantage is larger.

CLC-KSVD VS. exemplar: CLC-KSVD and exemplar dictionaries yield almost equal recognition performance at high SNR levels. As SNR decreases, the CLC-KSVD learned dictionaries start to outperform the exemplar dictionary. It is interesting to see that on test set B the advantage of CLC-KSVD over an exemplar dictionary is larger than on test set A. This indicates that the class likelihood consistent dictionary learning helps to increase the generalization power of our sparse coding based recognizer.

Including or not including noise exemplars in the sparse coding: On test set A, which has the same noise types as the exemplars, including noise exemplars in the sparse coding shows a marginal negative effect on the performance at the high SNR levels, while it leads to significant improvement at low SNR levels. This holds both for KSVD and CLC-KSVD dictionaries. On test set B, however, the effect of adding noise exemplars differs slightly for KSVD and CLC-KSVD. Using a KSVD learned dictionary, the best performance is obtained at all SNR levels if no noise exemplars are added. CLC-KSVD also shows the best performance at high SNR levels without adding noise exemplars. However, at the lowest SNR levels it appears to benefit from including noise exemplars: at SNR= $-5dB$ the difference (39.78-38.60%) is statistically significant while at SNR= 0dB (68.59-68.08%) it is not. Overall, the relative improvement achieved by including noise exemplars is larger when they represent matched noise types (i.e., test set A) than when they represent unmatched noise types (i.e., test set B). Moreover, both on test set A and B, the effect of adding noise exemplars is more visible when the dictionary is trained using clean speech.

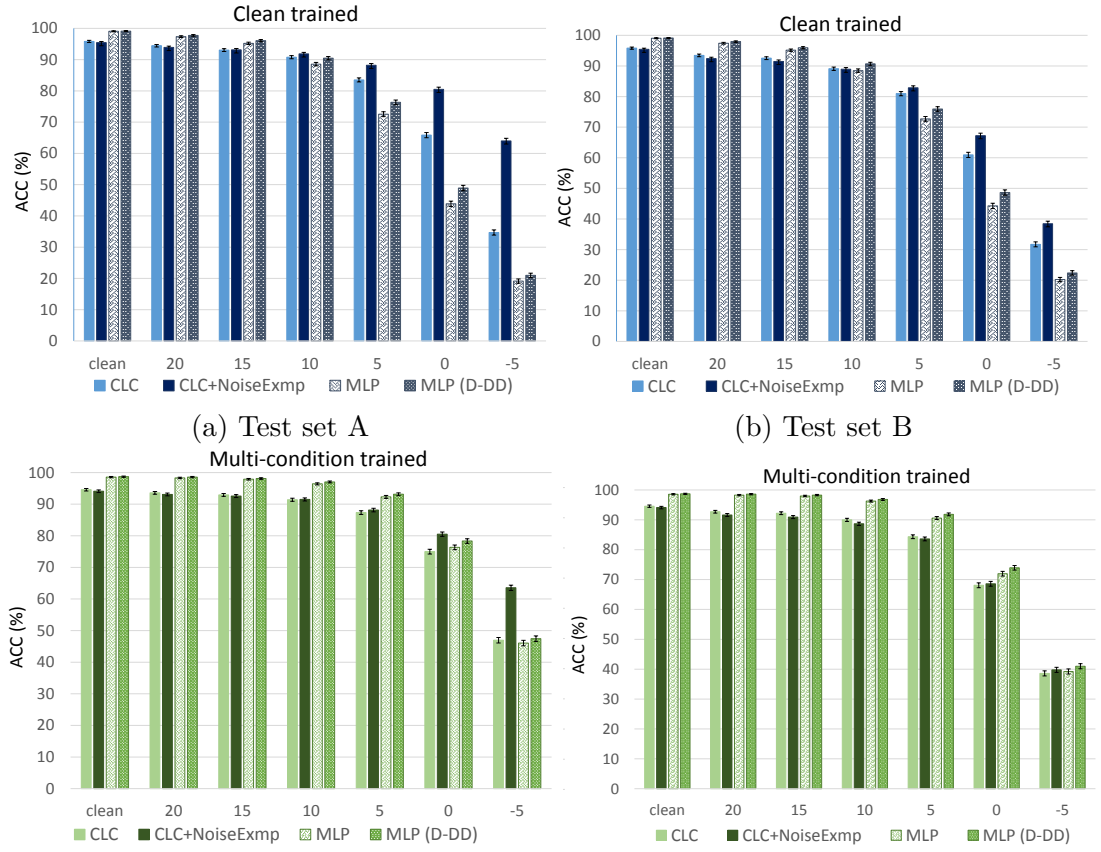
Clean or multi-condition training data: Both on test set A and B, multi-condition training is more effective in the absence of noise exemplars. On test set A, when noise exemplars are included, multi-condition training seems to have almost no effect at low SNR levels while it shows a minor, but statistically significant deterioration at high SNR levels. On test set B, however, the multi-condition training

seems to be more effective compared to including noise exemplars. This indicates that under mismatched noise conditions it is more beneficial to include noisified samples in training data rather than using noise exemplars explicitly in the sparse coding. While under matched noise conditions, including noise exemplars in the decoding is more effective compared to incorporating noisy data in training.

5.3.3.1 Comparison with MLP

In Chapter 3, we observed that a recognizer in which the posterior probabilities were estimated using MLP outperforms our exemplar based ASR system at high SNR levels. The experiments were performed on the 135-dimensional EMS features and their results are reported in reported in Table 3.2. To evaluate how much of the performance gap is narrowed by moving from exemplar based to learned dictionary based ASR, we repeated the MLP experiment using the 285-dimensional EMS features as our starting point. We trained four MLP networks. The first set of two was trained with the 285-dimensional EMS vectors. The other two were trained with EMS vectors enriched with their first and second derivative coefficients (Δ and $\Delta\Delta$). For each set of features, one of the MLPs was trained using clean and the other one using multi-condition training data. The first general observation we make is that the recognition accuracies obtained with the 285-EMS features are higher than with the 135-dimensional EMS features used in Chapter 3. The difference is around 1% in clean condition up to 11% at SNR= -5dB. This confirms the superiority of the enhanced EMS feature extraction designed in Chapter 4 over the initial design of Chapter 3. The enhanced features have higher dimensionality, however as discussed in Chapter4, the higher dimensionality of the feature space does not explain the performance improvement. Rather we believe the superior performance obtained using the enhanced design of EMS feature extraction is mainly related to the shape of the effective transfer function of the modulation filterbank (c.f. Section 4.3.3.2). Maximum accuracy obtained in the clean condition using MLPs on 285-dimensional EMS features is 99.13% which is close to the state-of-the-art performance on this data set and is higher than the best performance of MLP on PLP features reported in Table 3.2.

In Figure 5.3, the resulting recognition performance obtained with MLP is compared with the performance of the CLC-KSVD based recognizer. Figure 5.3a and 5.3b show the recognition accuracies obtained using the CLC-KSVD dictionary



(a) Test set A (b) Test set B

(c) Test set A (d)

FIGURE 5.3: Word recognition accuracies on AURORA-2 test data obtained with CLC-KSVD and MLP based recognizers. Error bars show the 5% confidence interval

and an MLP trained on clean speech. The results shown for CLC-KSVD corresponds to the sixth and seventh rows in Tables 5.4 and 5.5 in which the learned dictionary is used to compute the sparse codes both with and without including the noise exemplars. Similarly, Figures 5.3c and 5.3d are depicting the recognition accuracies obtained using the multi-condition trained CLC-KSVD dictionary (8th and 9th rows in Tables 5.4 and 5.5) and MLPs. The depicted MLP results from the clean and multi-condition trained networks are obtained using either EMS vectors or EMS+Δ + $\Delta\Delta$. Comparing the performance of MLPs with EMS and EMS+Δ + $\Delta\Delta$ shows that using the derivative coefficients are only useful at low SNR levels when the MLP is trained with clean data, otherwise the performance gain obtained by enriching the EMS vectors with derivative coefficients is marginal. This is in line with the observation we made from Table 3.2. On both test set A and B, the MLP outperforms the CLC-KSVD at SNR$\geq$ 15. As SNR decreases, the slope with which the performance of MLP system decreases

is steeper than that of CLC-KSVD. When only clean speech is used for training, both on test set A and B, CLC-KSVD outperforms MLP at SNR$\leq$ 5. In case of multi-condition training, CLC-KSVD combined with noise exemplars outperforms MLP at Low SNR levels in test set A, while on test set B the order is reversed; only at SNR=$-$5dB comparable performance levels are obtained.

5.4 Discussion

In this chapter we introduce a probabilistic classifier based on sparse coding representations in an over-complete basis. The over-complete basis is formed during a dictionary learning procedure. For this purpose we extended the K-SVD procedure for learning optimal reconstructive dictionaries in such a way that the Kullback-Leibler divergence between the class likelihood vectors produced by the classifier and the 'true' reference likelihood vectors is minimized together with the Euclidean distance between reconstructed and observed data vectors. The motivation to improve class likelihood estimation was that in many applications, the classification result is not the end product; rather, it is an input for later processing, such as speech decoder in an ASR system. Therefore, the final result does not only depend on the correctness of winning classes, but also on the likelihood scores of competitors.

We used a three-class sandbox task to prove that the algorithm underlying class-likelihood consistent KSVD (CLC-KSVD) operates as advertised, and that it results in a higher proportion of correctly classified observations and a closer approximation of the reference likelihoods. The knowledge of the true class likelihood vectors in the sandbox task makes it possible to compute the estimated upper bound of the classification accuracy. In most –if not all– real-world tasks the 'true' reference likelihood vectors are not known.

The simultaneous presence of two optimization criteria introduces the parameter λ that regulates the relative weight of reconstruction accuracy versus the KLD in CLC-KSVD training. In the sandbox task the optimal value was $\lambda = 0.8$, in the AURORA-2 task the optimum was at $\lambda = 0.002$, and in the TIMIT task the optimum was at $\lambda = 0.0008$. The range of the optimal value of λ for each task depends on the dimensionality of the feature vectors and the effective dimensionality of the reference likelihood vectors. These dimensions of the feature vectors

differed widely between the three tasks. The number of relevant entries in the class reference vectors depends on the way in which these vectors are defined in the two real-world tasks. For these reasons the optimal values of λ are difficult to compare between tasks, nor do they provide a direct indication of the relative weight assigned to the reconstruction accuracy relative to minimizing the KLD.

CLC-KSVD requires a definition of the reference class likelihood vectors, which are not available in the two speech tasks; therefore, these vectors must be created, in some way or other. Extensive experiments on the two speech tasks showed that for CLC-KSVD learning to be effective, it is necessary to use reference vectors with a fairly low entropy. In both tasks we ended up with reference likelihood vectors that have a substantially lower entropy than the entropy that would result from modelling the actual overlap of the classes in the observation spaces. Despite the discrepancy between the probability distribution in the actual observation space and the reference vectors both speech tasks showed that the frame-based performance indicators for CLC-KSVD are better than for KSVD.

Because the output of the classifiers serves as input for a back-end processor, it is tempting to use knowledge about the operation of that back-end in the creation of the reference likelihood vectors. Although the back-ends in the AURORA-2 and TIMIT tasks –both are essentially finite state acceptors– may seem to be quite similar, their interaction with the feeding probabilistic classifier is actually very different. In AURORA-2 the transducer consists of sequences of sets of 16 states that must be traversed from left to right, without skips. This structure benefits greatly from a classifier that produces an accurate approximation of the relative scores of the classes/states which rank highest in the output. This explains why the CLC-KSVD classifier with $\{\alpha = 0.4, \beta = 0.1\}$ in Table 5.2, for which %Correct@3 is maximal, also yields the highest word recognition accuracy, despite the fact that the other measures for this classifier are not the best. The fact that the back-end decoder cannot confuse the /aI/ in “five” with the /aI/ in “nine” without also confusing the preceding and following consonants was the motivation for not including confusions between similar sounds in different words in the definition of the reference vectors.

In TIMIT the interaction between the classifiers and the back-end is very different. The fact that we use single-state models, in combination with the fact that almost every phone can follow every other phone, implies that the back-end

cannot profit very much from improving the ranking of the highest-scoring classes. Therefore, we decided to base the definition of the reference likelihood vectors on the acoustic confusability. In AURORA-2 frame-based accuracies in the low thirties gave rise to word recognition accuracies in the mid nineties. In TIMIT frame-based accuracies in the mid thirties only resulted in phone recognition accuracy in the low fifties. Therefore, we must conclude that the capability of CLC-KSVD to improve %Correct and KLD at the level of individual observations is not guaranteed to improve the performance of each and every back-end. In future research we will investigate whether the phone recognition accuracy can be improved by using three-state phone models, which re-introduce some of the path constraints that proved so beneficial in the AURORA-2 task.

Sparse coding was not invented with classification as the primary goal. While CLC-KSVD learning adds a degree of supervision to the training process, it does not result in full-fledged discriminatory training. Therefore, it cannot be expected that classifiers based on sparse coding, even with the use of CLC learning, will always be the optimal choice in any application. Our previous experience with using sparse coding classifiers in speech recognition has shown that this type of classifiers is superior when the conditions in which the classifiers must be used are very different from the conditions under which they can be trained (Gemmeke et al., 2011b; Ahmadi et al., 2014). Noise-robust automatic speech recognition and noise-robust image classification in applications where only limited amounts of labelled training data are available, while the noise is extremely variable and unpredictable, are obvious examples of such situations. If train and test conditions are more similar, more conventional generative (GMM) and discriminative (Neural networks) classifiers are likely to be superior.

Our experiment on noisy speech recognition confirms that the speech recognition based on estimates obtained from sparse coding based classifiers are most effective in the presence of strong noise. While at high SNR levels MLP-based recognizers outperform any sparse coding based recognizer, when the SNR decreases, the advantage of sparse coding becomes visible. Our proposed CLC-KSVD approach shows competing performance with the exemplar approach at high SNR levels. As SNR level decreases, CLC-KSVD clearly outperforms the exemplar approach. We also observed that dictionary learning helps to increase the generalization power of the system and makes it more robust against noise mismatch conditions. Our experiments shows that our CLC-KSVD based classifier is the best choice in the

absence of any prior knowledge about noise where the only available option is to train the recognizer with clean speech.

5.5 Conclusion

In this chapter we introduced a mathematically tractable procedure for extending KSVD-based dictionary learning for creating classifiers based on sparse coding such that the Kullback-Leibler divergence between the classifier output and the true reference class likelihood is minimized at the same time that reconstruction accuracy is maximized. In a sandbox task, in which the true reference class likelihoods are known exactly the Class-Likelihood-Consistent learning procedure was shown to result in superior classification performance. Many applications that require a probabilistic classifier for providing input to a back-end decoder benefit more from an accurate approximation of class probability vectors than from maximizing the proportion of observations for which the winning class is correct. The effectiveness of CLC-KSVD learning was confirmed in the AURORA-2 connected digit recognition task, where the back-end benefits from a high ranking of the correct class, even if it is not the winner. The advantage of CLC-KSVD is more visible in noisy speech recognition when the SNR level approaches 0 dB. In our implementation of the TIMIT phone recognition task, in which accurate approximation of the likelihood vectors can hardly be used to advantage by the back-end, CLC-KSVD classifiers could not improve the performance of the complete system.

Chapter 6

General discussion and conclusions

In this thesis, we investigated to what extent the incorporation of various types of knowledge about human speech perception in noise can be exploited to improve the noise robustness of ASR systems. The thesis revolves around three main topics:

1. Feature enhancement using missing feature imputation;
2. Human inspired feature extraction;
3. Supervised dictionary learning for sparse classification of speech units.

Below we will formulate, per topic, the answers to the research questions that were put forward in Chapter 1. While summarizing the findings reported in the previous chapters, we will evaluate the completeness of these answers and discuss possible open issues for future research.

6.1 Feature enhancement using missing feature imputation

In Section 1.1 we raised the question: to what extent can we benefit from the underlying manifold of speech to reconstruct the missing parts of an incomplete

speech spectrogram? The underlying idea was that the acoustic variations in speech are limited because they are produced by an articulatory apparatus that cannot change the shape of the vocal tract in arbitrary ways. Therefore, we can assume that speech signals form a (nonlinear) manifold of much lower dimensionality than that of the embedding signal space (and, more importantly, that any deviations from the manifold are likely due to external noise sources). It has also been hypothesized that such a manifold also exists in the human brain as patterns of stable neural activities (Seung and Lee, 2000). To employ this knowledge in automatic speech recognition, we proposed to model this nonlinear manifold with MPPCA which is a mixture of linear sub-models (see Chapter 2). The dimensionality of these sub-models reflects the effective dimensionality of the manifold, and the optimum number of sub-models in the mixture shows how complex the manifold shape is in reality. We found the optimal dimensionality to be $q \approx 20$ which is in the same ballpark as the number of control parameters that are typically used in (articulatory and formant) speech synthesis systems for specifying the dynamic changes in the vocal tract geometry (Coker, 1976; Maeda, 1982; Klatt and Klatt, 1990). The optimal number of sub-models appeared to be in the range of $30 - 35$ which results in a much more compact model compared to other models (such as GMMs) used in MDT approaches that do not focus on the intrinsic structure of speech. It is worth noting that the optimal model dimension was the same both for small and medium vocabulary size data sets. This supports our underlying idea that the manifold captures the limited degrees of freedom of the articulatory apparatus: Irrespective of the size of the vocabulary, all speech signals can be situated on the manifold that is formed by all the possible variations in speech sounds of a particular language.

We showed that as long as there is a sufficient number of reliable measurements in the spectro-temporal representation of a speech signal, the position of (the nonlinear projection of) the signal in the latent space can be accurately located on the manifold. We also developed an inference procedure that uses the estimated location of the projection of the signal on the manifold to impute spectro-temporal values that are unlikely to pertain to speech but are more likely to be caused by external noise sources. The recognition accuracy on the reconstructed spectrograms improved significantly both in small and medium vocabulary noisy speech recognition tasks. Furthermore, our experimental results show that the recognition performance is relatively insensitive to acoustic variations introduced by external

sources, such as noise types and different microphones. Again, we interpret this as a confirmation that the manifold effectively captures the natural variation of speech in the acoustic space. Thanks to this property, imputation of missing data based on the manifold of speech does not involve any prior knowledge about noise. As a consequence, we observed comparable performance under matched and mismatched noise conditions. The very low dimensionality of the manifold makes it possible to find the location of the projection of (incomplete) input speech spectrograms on the manifold using a small number of reliable measurements. Therefore, even when the signal is highly corrupted by noise, the very few remaining reliable components are enough to achieve a reasonably accurate reconstruction. This is confirmed by the relatively high recognition accuracy obtained at very low SNR levels in conditions where we used oracle knowledge to determine reliable components. Unfortunately, this does not mean that manifold-based imputation is a panacea for noise robust ASR. It still requires estimating which spectro-temporal features really represent speech and can be considered reliable, a process which is inherently error-prone. Consequently, in real-life situations, the performance with an estimated mask is substantially reduced compared to oracle performance. This is a well-known problem that is shared by virtually all missing data imputation approaches. Although, in our results, this performance reduction is less compared to competing imputation methods, it still imposes a serious limit on the success of our MPPCA imputation in very noisy conditions. The gap between recognition performance using oracle and estimated mask indicates that there is room for improvement in determining reliable measurements in the noisy spectrogram. One obvious solution might be to rely on the development of more sophisticated mask estimation methods to improve mask estimation accuracy. Most classical mask estimation methods rely on a "blind" processing of frames to estimate the ratio of energy levels of speech and background noise (SNR); they don't involve any knowledge about the underlying speech signal. This local, context independent processing puts an upper limit to the success of such mask estimation methods. It is known that even small mask estimation errors can have a substantial impact on recognition performance (Gemmeke, 2010) and that often a correct identification of the spectrogram portions that pertain to speech can only be made if sufficient time context is available. Various approaches to incorporate context information in the mask estimation procedure have been proposed before, e.g., Van hamme (2004) and Barker et al. (2010). The reconstruction procedure used in our MPPCA imputation approach introduces an interesting alternative for making the

mask estimation more subtle and less blind. As a first step, it would be possible to replace the binary decision about the reliability of spectro-temporal measurements by probabilistic estimates (i.e., using probabilistic masks rather than binary masks, a solution that has been proven fruitful in competing missing data techniques. In addition, and most interestingly, since the MPPCA model reflects the intrinsic structure of the speech signal, it would be relatively straightforward to use the model also for estimating the probability that a spectro-temporal measurement pertains to speech or noise. By including the model of the speech signal in the mask estimation, it can help to reduce the chance of missing important content of the signal.

Thus, the idea of using the manifold model for mask estimation would allow to make the probabilistic mask estimation an integral part of the reconstruction procedure.

We showed that the manifold of speech can be effectively modeled using a mixture of probabilistic analyzers. We speculated that the dimensionality of sub-models ($q \approx 20$) is associated with the number of degrees of freedom in the human speech production system. Reasoning along similar lines it is tempting to assume that the number of sub-models (30-35) has a relation with the number of distinctive phones (39 according to Halberstadt and Glass (1997)). However, more in-depth investigations are needed to verify that such a relation between sub-models and phones or groups of phones indeed exists.

The potential of a manifold model to discriminate between speech and noise is not limited to missing data imputation. In fact, it opens up a wide range of other application areas where one would benefit from a reliable mechanism to separate speech variations from variations caused by external factors (e.g., speech enhancement, channel normalization and so on). The manifold model we proposed for speech can also be used to enhance our understanding of the structure of speech through an in-depth investigation of effects of MPPCA model parameters and hyper-parameters. It could reveal to what extent the complex shape of the manifold indeed reflects the articulatory constraints that help to avoid phonetically implausible/impossible speech confusions. Having a better insight in how distances in terms of speech production and perception are distributed along the manifold might make it possible to directly incorporate the manifold model in automatic speech recognition. For example one might use the position of speech signals on

the manifold as features to limit the search space for the ASR system. This would be particularly useful in noise robust ASR since it would allow to mimic the fact that human speech recognition in noise also seems to rely heavily on the most salient features only.

6.2 Human inspired feature extraction

In Section 1.2, we raised the question: Since modulation spectrum features appear to be a suitable basis for predicting intelligibility of speech in noise, to what extent are those features also an adequate starting point for a sparse coding based ASR system that is robust against noise? This research must be viewed against the background of our long term goal to move towards developing an ASR system that has a more human-like behavior. The model used for speech intelligibility prediction in Jørgensen and Dau (2011) aims to capture the properties of the human peripheral auditory system that are most relevant for intelligibility by computing a signal-to-noise ratio that is based on the modulations in the envelopes of an auditory filterbank. Despite the fact that the model provides the complete modulation spectrum of a speech signal, much of the details in this information remains unused since the computation of an SNR merely requires a long-term average of the output of the model. To what extent a full modulation spectrum could bring some of the desired noise robustness properties of humans to ASR was a largely unexplored territory.

In our first attempts to use the model from Jørgensen and Dau (2011) in an unchanged form for ASR, it soon became clear that the modulation filterbank had to be adapted to ensure that the modulation frequencies that are representative for speech were properly emphasized (and irrelevant modulations de-emphasized). Using knowledge from the literature, we modified the envelope modulation spectrum (EMS) feature extraction module accordingly. We tested its adequacy for ASR purposes by implementing an ASR system that can handle large and redundant input vectors. In our system the posterior probabilities of sub-word units were estimated using modulation spectrum features during an exemplar-based sparse coding procedure.

Despite the numerous sub-optimal design choices that our first ASR system (developed in Chapter 3) may have suffered from, we showed that the output of a

modulation spectrum analyzer is able to exploit the spectro-temporal continuity constraints that are typical for speech and which are a prerequisite for noise robust ASR. The preliminary experiments confirmed our intuition that a multiple resolution and redundant filterbank representation of speech may be an effective means to improve noise robustness. The resulting ASR system showed promising performance especially in very noisy conditions. The study also helped to identify two issues that served as motivation for the second phase of the study and also the study of Chapter 5:

1. Although the modulation processing was adapted for ASR purposes based on knowledge gathered from literature, we were not yet convinced that the exact way in which the different modulation frequencies were covered in our implementation were optimal for ASR purposes. Therefore, a systematic investigation with respect to the required range of modulation frequencies and the resolution with which the modulation frequency range is represented, was deemed in order.

2. We realized that there is a fundamental distinction between *sparse coding*, where the task is to find the optimal representation of an unknown observation in a very large dimensional space and *sparse classification* where the task is to obtain the best possible estimation of posterior probability that an unknown observation belongs to a specific class. We also observed that MLP classifiers outperform our exemplar-based system in high SNR conditions. This indicates that our ASR system can benefit from including proper supervised learning in which the accuracy in estimating class posterior probabilities is taken into account in the objective function (in addition to or as a alternative for reconstruction error).

The first issue was addressed in Chapter 4. Although one might prefer an optimal configuration of the filterbank based on perceptual relevance of modulation frequencies, in practice, this is infeasible because the required knowledge is incomplete. A closed-form, mathematical solution is also impossible. It would require optimization of a cost function derived from the mathematical model of our ASR system. However, since the ASR system is a nonlinear multi-module system, it is impossible to define an analytical relation between the design parameters of the filterbank and the recognition performance. Therefore, we resorted to a systematic,

empirical investigation of the configuration of the modulation frequency analyser consisting of a bank of one low-pass and a set of band-pass filters (referred to in short as a modulation filterbank).

We found that for maximizing ASR performance we had to use a higher frequency resolution than was found to be adequate for intelligibility prediction in Jørgensen and Dau (2011). [1] We also found that there is no unique configuration of the modulation filterbank that is optimal for all SNR levels and all noise types. For example at the lowest SNR levels, our ASR system benefits from a high resolution in the lowest modulation frequencies indicating that, in noisy circumstances, it is beneficial to rely more heavily on long-term time context. This is a plausible finding because it allows the system, analogous to humans, to profit from more glimpses of the speech signal that stick out of the noise (Cooke, 2006). Although from an engineering point of view, it is satisfying that a reasonable compromise can be found leading to near-optimal performance in most conditions, it is intriguing to note that an SNR dependent processing is almost mandatory to achieve the best results possible. At this point in time our ASR system is far too crude to be considered as an ecologically valid model of human speech perception. Moreover, the poorly understood nonlinear interactions between the various modules in our ASR system do not warrant any far reaching conclusions regarding striking commonalities with human behaviour. Yet, it is worthwhile to note that also humans seem to focus on different signal properties dependent on the noise level (e.g. in Shi et al. (2006) it is found that phase information is relevant at low SNR-levels only).

After optimization of the modulation filterbank, the resulting ASR system outperformed traditional MFCC-based recognizers at low SNR levels, at least in matched conditions where the same noise types were present as in the training set. Our ASR system that employs single-frame EMS features also shows superior performance compared to a similar exemplar-based SC system that uses 30-frame Mel-band features. This warrants the conclusion that splitting up the time envelopes of different auditory filterbands in multiple modulation frequency bands helps to make ASR more noise robust. We believe that mimicking the modulation filter analysis from the human auditory system makes it easier to determine the

[1] At this point in time we don't have any information to what extent speech intelligibility prediction would also benefit from more resolution in the modulation frequency domain.

similarity between speech signals that are corrupted by various levels of noise because one can more easily disregard all energy fluctuations in modulation bands that are unlikely to be caused by speech production. Yet, the performance of our ASR system is still far from human-like. Like any other exemplar-based ASR system that uses a noise dictionary, our system does not generalize to mismatched conditions (unseen noise types). Moreover, in its current implementation, the system has sub-standard performance in quiet conditions, both compared to HSR and to traditional MFCC-based ASR systems.

The absolute performance levels of our EMS+SC system cannot rival with human recognition performance. Nevertheless, our system mimics human speech recognition on a semantic free task better than traditional MFCC+GMM systems. As shown in Section 4.4.2, the SNR at which the recognition accuracy is 50% (speech reception threshold) of our system is closer to that of HSR. The easy and difficult noise types are also shared between our system and HSR while traditional systems often have the lowest performance with noise types that are not the most challenging for HSR. The only discrepancy is that HSR exhibits lower accuracy for long utterances while our system does not show this effect because it does not include any simulation of working memory. It seems that using EMS features is a successful first step towards making the behaviour more human-like. A detailed comparison of confusions made by an ASR system and human listeners on the same recognition task might prove a powerful instrument to reveal the strengths and weaknesses of the current implementation.

We started our enterprise entirely from an auditory modelling perspective. However, we found there is an interesting link between our EMS features and Mel-band features in more traditional ASR approaches. The EMS feature vectors can be considered as a combination of smoothed static filterbank features and a set of accompanying dynamic delta features that are smoothed to a different degree and therefore also span different time context. Since our experiments showed that the multiple dynamic channels are helpful in dealing with noise, we might conclude that not only our sparse coding based system, but in fact any classical ASR system would benefit from a frontend in which static and dynamic features are represented in a multi-resolution fashion.

As explained previously, there are a multitude of issues to be resolved before our ASR system can be considered to mimic human-like behaviour both in terms of

absolute performance and in terms of recognition errors. During our experiments we touched upon a number of phenomena that were considered interesting in their own right but for which there was not sufficient time to investigate in any depth. Below we present a (non-exhaustive) list of future work in unexplored areas that we believe might contribute to improve the system:

- The adaptation/compression network in Dau's model (Dau et al., 1996) that is absent in the model used for intelligibility prediction seems to be required to better detect word onset and offsets. Preliminary experiments (not reported in this thesis) using static compression on clean speech confirmed this. For noisy speech, however, static compression appeared not to be helpful and a more powerful adaptive dynamic compression mechanism would need to be developed to better emphasize onset and offset events in the various modulation frequency channels.

- Above we mentioned the analogy between the modulation features extracted by modulation filters (with quality factor $Q = 1$) and the traditional Δ features. Since we observed that having multiple modulation bandfilters in parallel increased noise robustness, it would be interesting to see if extending the EMS feature space by multi-range static features would have a similar effect. Furthermore, since we know from traditional ASR systems that acceleration ($\Delta\Delta-$)coefficients typically improve the absolute recognition performance, a further extension of the EMS feature vector with the output of $Q = 2$ filters would also seem logical.

- The Euclidean distance we were forced to use because of having bipolar data does not necessarily represent a distance that matches the perceptual distance a human would experience between an actual observation and stored exemplars. We counteracted some of the undesired effects of the Euclidean distance using some empirically designed equalization and normalization of features. It remains an open question, however, whether it is possible to transform the feature vectors to a space in which the Euclidean distance gauges the distances of objects in a certain neighbourhood in a similar way as a human observer would.

- The performance of our system, as other exemplar based systems, is inferior at high SNR levels. Future research will have to reveal what causes the

confusions in exemplar based systems at high SNR levels. We observed that using MLP neural networks on EMS features to estimate posterior probability of subword units leads to superior recognition performance, at high SNR levels, compared to sparse coding. At low SNR levels, however, the sparse coding shows significant superiority. It remains a question whether more complex neural networks such as DNNs will be able to benefit from the salient information in the EMS features that sparse coding benefits from at low SNR levels and lead to higher accuracy compared to MLPs.

6.3 Supervised dictionary learning for sparse classification of speech units

In our study on human inspired features (see Chapter 3), we observed that the exemplar based ASR system suffers from low performance at high SNR levels compared to the learning based recognizers such as MLP. This instigated us to include learning in the construction of the dictionary for sparse coding. Since the main objective in our posterior estimation procedure is sparse classification, we asked the question: To what extent can we improve the accuracy of the estimated posterior probabilities of speech units by customizing the learning procedure in such a way that the probabilistic classification accuracy is also considered as an objective during learning? To address this question, we proposed class likelihood consistent dictionary learning. This newly proposed learning procedure not only minimizes the reconstruction error of sparse coding, but is additionally constrained in such a way that the sparse codes also provide an accurate estimation of the class membership probabilities for each observation. We realized this by developing an objective function which is a regularized combination of the reconstruction error and an upper bound for the Kullback-Leibler divergence between the estimated and true class probability vectors.

On sets of synthetic data as well as on speech data, we were able to show that the proposed method yields superior performance both in terms of classification accuracy and in terms of accuracy with which class likelihoods are estimated. In the test on synthetic data created by some pre-defined model, the available distribution function of the data provided a reference to evaluate the classification

performance. On the speech data, on the other hand, the classification performance was judged by means of the eventual word/phone recognition accuracy. We observed that the performance of the sparse coding system can be significantly improved by using a learned dictionary instead of an exemplar dictionary of the same size. Among the learned dictionaries, the best recognition performance was obtained using the class likelihood consistent learning. The superior performance of the CLC learning is caused not only by the fact that the learning is supervised, but also (and most importantly) by the way in which the supervision is applied. Imposing class likelihood consistency provides a more reasonable set of probable classes for each input feature vector and this is beneficial when a search algorithm must find the most probable sequence of classes (sub-words or phones).

The class likelihood consistent learning appeared to be successful in decreasing the KL-divergence between estimated and true likelihood vectors. However, it remains a challenge to find suitable reference true likelihood vectors for real world tasks. Although for most classification tasks, labelled data is available, rarely golden class likelihood vectors are provided. However, as discussed in Chapter 5, it is often possible to derive reasonable reference estimates for true likelihoods by analysing the geometric distribution of labelled data which reflects the physical overlap between classes. Alternatively, the true likelihoods can also be defined based on the preferred class confusions by the backend that has to process the resulting estimated probabilities. In our research, we employed both strategies and observed that in both cases the proposed probabilistic classifier benefits from imposing likelihood consistency using these reference vectors.

We then continued this study by the question: To what extent does posterior probability estimation based on learned dictionaries increase noise robustness in the sparse coding based ASR system? Using the dictionary learned with the proposed method to estimate the posterior probability of HMM states in AURORA-2 noisy test data we compared the recognition results with the best results we could get with exemplar dictionaries containing both speech and noise exemplars. We observed that the class likelihood consistent learned dictionary combined with an exemplar noise dictionary outperforms the speech+noise exemplar dictionaries for SNR<10 dB both in noise matched and mismatched conditions. This result indicates that the variation in multiple exemplars pertaining to the same underlying linguistic message is successfully captured by the atoms of the learned dictionary and yields more robustness against noise. We also observed that under the clean

training condition, our learned dictionary outperforms an MLP system at low SNR levels both on matched and mismatched conditions. We hence conclude that in the absence of noise samples to train the classifier with or to form a noise exemplar dictionary from, our proposed learning method will provide a solution which, compared to MLP or an exemplar approach, is more noise robust. Meanwhile, it is reasonable to expect that adding a learned noise dictionary will further improve the performance of our system and its generalization power. However, including noise information in the learned-dictionary-based SC recognizer requires more thorough investigation. This opens up a new area to explore that is not covered by this thesis.

6.4 Conclusions and future outlook

In this thesis we conducted three studies in which we tried to improve the noise robustness of ASR systems by incorporating knowledge about human speech perception in noise. Each of the three studies were focused on very different aspects and had specific advantages and disadvantages. Hence, it is not straightforward to draw a single overall conclusion from the whole project. However, we believe the insights gained by each of the three studies can be combined to sketch a promising future.

We proposed a novel feature extraction inspired by models of the human auditory system which provided a high dimensionality and redundant feature space. The novel features combined with an exemplar based sparse coding ASR backend led to competitively high performance at low SNR levels. Although the performance at high SNR levels was limited because of the shortcomings of the SC backend, the competing performance we got using MLP on our proposed features reveals the distinctive power of the feature space. These observations all indicate that although sparse coding is an effective tool in separating speech and noise, it is not a sufficiently sensitive tool for distinguishing speech units. That is why our sparse coding system only outperforms other backends such as GMM-HMM and MLP recognizers at low SNR levels. Although we succeeded to improve the performance of the sparse coding backend by replacing the exemplar dictionary with a learned one and introducing likelihood consistent learning, this approach was again more effective at low SNR levels. We hence conclude that to maximally benefit from

the richness of the modulation feature space at all SNR levels, it is necessary to combine systems that are able to discriminate speech and noise with systems that have sufficient discriminative power to distinguish speech sounds.

Combining the insights gained in the three studies may give further clues as how to innovate future ASR system configurations. In the first study we focused on the intrinsic structure of the speech manifold instigated by the finding that this manifold is also observed in the human perception system to handle speech variations. Manifold learning hinges on processing in a very low dimensional nonlinear space, as opposed to a sparse coding approach, which processes the feature vectors in an over-complete space spanned by dictionary exemplars. We showed that the compact and low-dimensional manifold model can be employed together with the reliable parts of the spectro-temporal features of speech to compensate the effect of noise. The feature space we obtained in our second study is of a much higher dimensionality and is also much more redundant compared to the spectro-temporal features that were used in the manifold approach. As indicated by our sparse coding experiments, in such a highly redundant feature space, there is a higher chance of finding feature elements that are exclusively representative of speech (and not of noise). This makes it interesting to investigate whether manifold learning can be more effective if it is performed on the proposed EMS features (either to impute missing data or to perform recognition). Since the manifold model refers to the underlying structure of speech, we do not expect a very different model to emerge but the mapping from feature space to the manifold is likely to be different. Having such a mapping would make it possible to benefit from the speech dominated feature elements available in the EMS feature vectors and to identify noisy speech based on the resulting position on the manifold. Consequently, we expect the resulting position of noisy speech on the manifold to be more accurate (compared to, for instance, classical Mel features) and superior recognition performance both in clean and noisy circumstances to become within reach.

Up to now we investigated the manifold model and learning-based sparse coding separately. Also, we have no insight yet into how the manifold model looks like when applied to the proposed EMS feature space. Nevertheless, it is tempting to conjecture that combining the approaches allows to profit from the best of two worlds. We envision a recognition system in which EMS feature vectors go through two parallel processing lines. The first processing line would focus on identifying the speech segments based on their corresponding location on the low-dimensional

manifold of speech, thus allowing to decode the noisy speech or to impute the missing parts and decode the reconstructed speech. The second processing line would then be concerned with estimating posterior probabilities of sub-word units using the sparse codes computed in an over-complete space provided by a supervised learned dictionary. Building on previous experiences (Ellis, 2000; Hennansky et al., 1996; Kirchhoff and Bilmes, 2000; Li et al., 2002; Singh et al., 2001), it is reasonable to expect that fusing the outputs of the two, quite different processing lines would lead to a more accurate and more noise robust speech recognition.

Appendix A

MAP estimation of MPPCA model parameters in the missing data reconstruction phase

Having the posterior distribution of unknown variables $\gamma = \{\vec{S}, \vec{x}, \vec{t}_u\}$, the MAP estimation is

$$\hat{\vec{\gamma}} = \underset{\vec{\gamma}}{\text{argmax}} \left\{ P(\vec{\gamma}|\vec{t}_r) \right\}. \tag{A.1}$$

The posterior distribution can be written as

$$P(\vec{\gamma}|\vec{t}_r) = P(\vec{S}, \vec{x}, \vec{t}_u|\vec{t}_r) = P(\vec{S}, \vec{x}, \vec{t}_u, \vec{t}_r)/P(\vec{t}_r). \tag{A.2}$$

The numerator of the above fraction can be rewritten as

$$P(\vec{S}, \vec{x}, \vec{t}_u, \vec{t}_r) = P(\vec{S}, \vec{x}, \vec{t}) = P(\vec{t}|\vec{x}, \vec{S})P(\vec{x})P(\vec{S}). \tag{A.3}$$

Using equation (2.1) in the text and given the Gaussian distribution of $\vec{\boldsymbol{\varepsilon}}$, the posterior probability of the observation vector $\vec{t}$ given the latent variables can be written as:

$$P(\vec{t}|\vec{x},\vec{S}) = \mathcal{N}(\vec{t}|\vec{W}_m\vec{x} + \vec{\mu}, \tau^{-1}\vec{I}_d). \tag{A.4}$$

The distributions of $\vec{x}$ and $\vec{S}$ mentioned in training are valid in reconstruction too, i.e.,

$$P(\vec{x}) = \mathcal{N}(\vec{x}|0, \vec{I}_q). \tag{A.5}$$

$$P(\vec{S}|\vec{\pi}) = \prod_{m=1}^{M} \pi_m^{Sm}. \tag{A.6}$$

Hence eq. (A.2) can be written as

$$P(\vec{S},\vec{x},\vec{t}_u|\vec{t}_r) = \frac{1}{P(\vec{t}_r)} \prod_{m=1}^{M} \left\{ \begin{aligned} &\pi_m \mathcal{N}(\vec{x}|0,\vec{I}_q) \\ &\quad .\mathcal{N}(\vec{t}_r|\vec{W}_{m,r}\vec{x} + \vec{\mu}_{m,r}, \tau^{-1}\vec{I}_{d_r}) \\ &\quad .\mathcal{N}(\vec{t}_u|\vec{W}_{m,u}\vec{x} + \vec{\mu}_{m,u}, \tau^{-1}\vec{I}_{d_u}) \end{aligned} \right\}^{\vec{S}_m}, \tag{A.7}$$

where $\vec{S}_m$ is the m^{th} entry of $\vec{S}$ that can be either 1 or 0. As can be seen from eq. (A.7), $\vec{t}_u$ only appears in the last term, therefore its MAP estimation will be as follows:

$$\hat{\vec{t}}_u = \vec{W}_{\hat{m},u}\hat{\vec{x}} + \vec{\mu}_{\hat{m},u}. \tag{A.8}$$

Here $\hat{m}$ indicates the active component in $\hat{\vec{S}}$. Substituting $\hat{\vec{t}}_u$ in $P(\vec{S},\vec{x},\vec{t}_u|\vec{t}_r)$, the last term in (A.7) will be the constant value $(\tau/2\pi)^{\frac{d_u}{2}}$. In the remaining two terms, we multiply two normal distributions, which, using the hints given in Appendix B of Bishop (2006), can be simplified as

$$\mathcal{N}(0,\vec{I}_q)\mathcal{N}(\vec{t}_r|\vec{W}_{m,r}\vec{x} + \vec{\mu}_{m,r}, \tau^{-1}\vec{I}_{d_r}) \propto K_m \times \mathcal{N}(\vec{x}|\vec{\eta}_m, \vec{\Sigma}_m), \tag{A.9}$$

where

$$K_m = \exp\left(-(\pi/2)(\vec{t}_r - \vec{\boldsymbol{\mu}}_{m,r})^T \boldsymbol{\Lambda}_m (\vec{t}_r - \vec{\boldsymbol{\mu}}_{m,r})\right), \tag{A.10}$$

$$\boldsymbol{\Lambda}_m = (\vec{I} - \tau \vec{W}_{m,r} \vec{\Sigma}_m \vec{W}_{m,r}^T), \tag{A.11}$$

$$\vec{\Sigma}_m^{-1} = \vec{I}_q + \tau \vec{W}_{m,r}^T \vec{W}_{m,r}, \tag{A.12}$$

$$\eta_m = \tau \vec{\Sigma}_m \vec{W}_{m,r}^T (\vec{t}_r - \vec{\boldsymbol{\mu}}_{m,r}). \tag{A.13}$$

Accordingly, the MAP estimation of $\vec{x}$ will be the mean of the resulting normal distribution:

$$\hat{\vec{x}} = \vec{\boldsymbol{\eta}}_{\hat{m}}. \tag{A.14}$$

To estimate $\hat{\vec{S}}$, only the estimation of $\hat{m}$ is needed, since all entries of $\hat{\vec{S}}$ are zero except the $\hat{m}^{th}$ entry. Substituting $\vec{x}$ by $\hat{\vec{x}}$ in $K_m \times \mathcal{N}(\vec{x}|\vec{\eta}_m, \vec{\Sigma}_m)$, eq. (A.9) reduces to K_m multiplied by a constant. Hence, using eq. (A.7), the MAP estimation for $\hat{m}$ is obtained:

$$\hat{m} = \underset{m}{\operatorname{argmax}} \left\{\log(\pi_m) - d_m\right\}, \tag{A.15}$$

where d_m is a Mahalanobis distance

$$d_m = (\tau/2) \left(\vec{t}_r - \vec{\boldsymbol{\mu}}_{m,r}\right)^T \vec{\Lambda}_m \left(\vec{t}_r - \vec{\boldsymbol{\mu}}_{m,r}\right). \tag{A.16}$$

Bibliography

Acero, A. (1993). *Acoustic And Environmental Robustness In Automatic Speech Recognition*. Kluwer Academic Publishers, Boston.

Aharon, M., Elad, M., and Bruckstein, A. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322.

Aharon, M., Elad, M., and Bruckstein, A. M. (2005). K-SVD and its non-negative variant for dictionary design. In *Proceedings of Optics & Photonics 2005*, volume 5914, pages 11.1–11.13. International Society for Optics and Photonics.

Ahmadi, S., Ahadi, S. M., Cranen, B., and Boves, L. (2014). Sparse coding of the modulation spectrum for noise-robust automatic speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):1–20.

Arai, T., Pavel, M., Hermansky, H., and Avendano, C. (1999). Syllable intelligibility for temporally filtered LPC cepstral trajectories. *The Journal of the Acoustical Society of America*, 105(5):783–791.

Baby, D. and Van hamme, H. (2015). Investigating modulation spectrogram features for deep neural network-based automatic speech recognition. In *Proceedings of Interspeech*, pages 2479–2483, Dresden, Germany.

Bacon, S. P. and Viemeister, N. F. (1985). Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners. *International Journal of Audiology*, 24(2):117–134.

Baraniuk, R. and Wakin, M. (2009). Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77.

Barker, J., Cooke, M., and Green, P. (2001). Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. In *Proceedings of Eurospeech*, pages 213–216.

Barker, J., Ma, N., Coy, A., and Cooke, M. (2010). Speech fragment decoding techniques for simultaneous speaker identification and speech recognition. *Computer Speech & Language*, 24(1):94–111.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., et al. (2007). Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19:153–160.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Bishop, C. M. and Winn, J. (2000). Non-linear Bayesian image modeling. In *Proceedings of sixth European Conference on Computer Vision*, volume 1, pages 3–17.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, Speech and Signal processing*, 27(2):113–120.

Boureau, Y.-L., Bach, F., LeCun, Y., and Ponce, J. (2010). Learning mid-level features for recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2559–2566. IEEE.

Bourlard, H. (1999). Non-stationary multi-channel (multi-stream) processing towards robust and adaptive ASR. In *Proceedings of ESCA Workshop Robust Methods Speech Recognition in Adverse Conditions*, pages 1–10.

Bourlard, H., Dupont, S., Hermansky, H., and Morgan, N. (1996a). Towards subband-based speech recognition. In *Proceedings of EUSIPCO*, pages 1579–1582.

Bourlard, H., Hermansky, H., and Morgan, N. (1996b). Towards increasing speech recognition error rates. *Speech Commununication*, 18:205–231.

Bourlard, H. A. and Morgan, N. (2012). *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media.

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, Mass.

Buesing, L., Bill, J., Nessler, B., and Maass, W. (2011). Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, 7(12).

Candes, E. J. and Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30.

Cerisara, C. and Fohr, D. (2001). Multi-band automatic speech recognition. *Computer Speech and Language*, 15:151–174.

Chang, H.-A. and Glass, J. R. (2007). Hierarchical large-margin gaussian mixture models for phonetic classification. In *Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 272–277. IEEE.

Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D., and Carin, L. (2010). Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Transactions on Signal Processing*, 58(12):6140–6155.

Chen, Y.-C., Patel, V., Pillai, J., Chellappa, R., and Phillips, P. (2013). Dictionary learning from ambiguously labeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CPVR 2013)*, pages 353–360.

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906.

Choi, J., Cho, H., Kwac, J., and Davis, L. (2014). Toward sparse coding on cosine distance. In *Proceedings of 22nd International Conference on Pattern Recognition (ICPR)*, pages 4423–4428.

Coker, C. H. (1976). A model of articulatory dynamics and control. *Proceedings of the IEEE*, 64(4):452–460.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3):1562–1573.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2001). Robust automatic speech recognition with missing and uncertain acoustic data. *Speech Communication*, 34:267–285.

Cooke, M. P., Morris, A., and Green, P. D. (1997). Missing data techniques for robust speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-97)*, volume 2, pages 863 – 866.

Cui, X. and Gong, Y. (2007). A study of variable-parameter gaussian mixture hidden markov modeling for noisy speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1366–1376.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Mit Press.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997a). Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America*, 102(5):2892–2905.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997b). Modeling auditory processing of amplitude modulation. ii. spectral and temporal integration. *The Journal of the Acoustical Society of America*, 102(5):2906–2919.

Dau, T., Püschel, D., and Kohlrausch, A. (1996). A quantitative model of the "effective" signal processing in the auditory system. i. model structure. *The Journal of the Acoustical Society of America*, 99(6):3615–3622.

De Wachter, M., Matton, M., Demuynck, K., Wambacq, P., Cools, R., and Van Compernolle, D. (2007). Template-based continuous speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1377–1390.

Demuynck, K., Garcia, O., and Van Compernolle, D. (2004). Synthesizing speech from speech recognition parameters. In *Proceedings of Interspeech*, volume 2, pages 945–948, Jeju Island, Korea.

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *The Journal of the Acoustical Society of America*, 97(1):585–592.

Drullman, R., Festen, J. M., and Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, 95:1053–1064.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Elad, M. and Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745.

Elhilali, M., Chi, T., and Shamma, S. A. (2003). A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech communication*, 41(2):331–348.

Ellis, D. P. (2000). Stream combination before and/or after the acoustic model. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00)*, volume 3, pages 1635–1638. Citeseer.

Ewert, S. D. and Dau, T. (2000). Characterizing frequency selectivity for envelope fluctuations. *The Journal of the Acoustical Society of America*, 108(3):1181–1196.

Fazel, A. and Chakrabartty, S. (2012). Sparse auditory reproducing kernel (SPARK) features for noise-robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1362–1371.

Festen, J. M. and Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, 88(4):1725–1736.

Fletcher, H. (1940). Auditory patterns. *Reviews of modern physics*, 12(1):47.

Fletcher, H. (1953). *Speech and Hearing in Communication*. Krieger, New York.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):254–272.

Gandhi, M. and Jacob, J. (1998). Natural number recognition using MCE trained inter-word context dependent acoustic models. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-98)*, volume 1, pages 457–460.

Geiger, J. T., Weninger, F., Hurmalainen, A., Gemmeke, J. F., Wöllmer, M., Schuller, B., Rigoll, G., and Virtanen, T. (2013). The TUM+ TUT+ KUL approach to the 2nd CHiME challenge: Multi-stream ASR exploiting BLSTM networks and sparse NMF. *Proceedings of CHiME*, pages 25–30.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Gemmeke, J. (2010). *Noise Robust ASR: Missing data techniques and beyond*. PhD thesis, Radboud University, Nijmegen.

Gemmeke, J. F., Cranen, B., and Remes, U. (2011a). Sparse imputation for large vocabulary noise robust ASR. *Computer Speech and Language*, 25(2):462–479.

Gemmeke, J. F., Van hamme, H., Cranen, B., and Boves, L. (2010). Compressive sensing for missing data imputation in noise robust speech recognition. *IEEE Journal of Selected Topics Signal Processing*, 4(2):272–287.

Gemmeke, J. F., Virtanen, T., and Hurmalainen, A. (2011b). Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Transactions on Audio, Speech and Language processing*, 19(7):2067–2080.

Ghitza, O. (1994). Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(1):115–132.

Ghosh, P. and Narayanan, S. (2011). Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America*, 130(4):251–257.

Glasberg, B. and Moore, B. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–138.

Glickman, O., Dagan, I., and Koppel, M. (2005). A probabilistic classification approach for lexical textual entailment. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 1050–1055.

Goldinger, S. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological review*, 105(2):251–279.

González, J., Peinado, A. M., Gómez, A. M., Ma, N., Barker, J., et al. (2012). Combining missing-data reconstruction and uncertainty decoding for robust speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'12)*, pages 4693–4696. IEEE.

Gonzalez, J. A., Peinado, A. M., Ma, N., Gomez, A. M., and Barker, J. (2013). MMSE-based missing-feature reconstruction with temporal modeling for robust speech recognition. *IEEE Transactions on Audio, Speech, Language Processing*, 21:624–635.

Grossberg, S. and Kazerounian, S. (2011). Laminar cortical dynamics of conscious speech perception: Neural model of phonemic restoration using subsequent context in noise. *The Journal of the Acoustical Society of America*, 130(1):440–460.

Haeb-Umbach, R. and Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-92)*, volume 1, pages 13–16. IEEE.

Halberstadt, A. K. and Glass, J. R. (1997). Heterogeneous acoustic measurements for phonetic classification. In *Proceedings of Eurospeech*, pages 401–404.

Han, Y., de Veth, J., and Boves, L. (2007). Trajectory clustering for solving the trajectory folding problem in automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1425 – 1434.

Hennansky, H., Tibrewala, S., and Pavel, M. (1996). Towards ASR on partially corrupted speech. In *Proceedings of Fourth International Conference on Spoken Language (ICSLP 96)*, volume 1, pages 462–465. IEEE.

Henry, M. J., Herrmann, B., and Obleser, J. (2015). Selective attention to temporal features on nested time scales. *Cerebral Cortex*, 25(2):450–459.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752.

Hermansky, H. (1997). The modulation spectrum in the automatic recognition of speech. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 140–147, Santa Barbara.

Hermansky, H. (2011). Speech recognition from spectral dynamics. *Sadhana (Indian Academy of Sciences)*, 36(5):729–744.

Hermansky, H. (2013). Multistream recognition of speech: Dealing with unknown unknowns. *Proceedings of the IEEE*, 101(5):1076–1088.

Hermansky, H. and Fousek, P. (2005a). Multi-resolution RASTA filtering for TANDEM-based ASR. In *Proceedings of Interspeech*, pages 361–364, Lisbon.

Hermansky, H. and Fousek, P. (2005b). Multi-resolution rasta filtering for TANDEM-based ASR. In *Proceedings of International Conference of Spoken Language Processing*, pages 361–364.

Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. (1991). Compensation for the effect of the communication channel in auditory-like analysis of speech RASTA-PLP. In *Proceedings of Second European Conference on Speech Communication and Technology*, pages 1367–1370.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hirsch, H. and Pearce, D. (2006). Applying the advanced ETSI frontend to the Aurora-2 task. *in version*, 1.1.

Hirsch, H. G. and Pearce, D. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings of ISCA Workshop ASR2000, Automatic Speech Recognition: Challenges for the Next Millennium*, pages 29–32, Paris, France.

Holmes, J. and Holmes, W. (2001). *Speech Synthesis and Recognition*. Taylor and Francis, London and New York, 2 edition.

Horton, P. and Nakai, K. (1996). A probabilistic classification system for predicting the cellular localization sites of proteins. In *Proceeding of the Fourth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, volume 4, pages 109–115, St Louis, MO, USA. AAAI press.

Houtgast, T. (1989). Frequency selectivity in amplitude-modulation detection. *The Journal of the Acoustical Society of America*, 85(4):1676–1680.

Houtgast, T. and Steeneken, H. J. M. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America*, 77:1069–1077.

Hromádka, T., DeWeese, M. R., and Zador, A. M. (2008). Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biol*, 6(1):e16.

Huang, H., Liu, Y., Gemmeke, J., ten Bosch, L., Cranen, B., and Boves, L. (2011). Globality-locality consistent discriminant analysis for phone classification. In *Proceedings of Interspeech*, pages 1253–1256.

Huang, H., Liu, Y., ten Bosch, L., Cranen, B., and Boves, L. (2016). Locally learning heterogeneous manifolds for phonetic classification. *Computer Speech & Language*, 38:28 – 45.

Huang, K. and Aviyente, S. (2006). Sparse representation for signal classification. In *Proceedings of Advances in neural information processing systems (NIPS)*, pages 609–616.

Huang, X., Acero, A., Hon, H.-W., and Foreword By-Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.

Hurmalainen, A., Mahkonen, K., Gemmeke, J. F., and Virtanen, T. (2011a). Exemplar-based recognition of speech in highly variable noise. In *Proceedings of CHiME*, pages 1–5.

Hurmalainen, A., Mahkonen, K., Gemmeke, J. F., and Virtanen, T. (2011b). Exemplar-based recognition of speech in highly variable noise. In *Proceedings of International Workshop on Machine Listening in Multisource Environments*, Florence.

Jansen, A. and Niyogi, P. (2013). Intrinsic spectral analysis. *IEEE Transactions on Signal Processing*, 61(7):1698–1710.

Jiang, W., Zhang, Z., Li, F., Zhang, L., Zhao, M., and Jin, X. (2016). Joint label consistent dictionary learning and adaptive label prediction for semisupervised machine fault classification. *IEEE Transactions on Industrial Informatics*, 12(1):248–256.

Jiang, Z., Lin, Z., and Davis, L. S. (2013). Label consistent K-SVD: Learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc.

Johnson, D., Ellis, D., Oei, C., Wooters, C., Faerber, P., Morgan, N., and Asanovic, K. (2004). ICSI quicknet software package. http://www.icsi.berkeley.edu/Speech/qn.html. Online; accessed 1-June-2013.

Jørgensen, S. and Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *The Journal of the Acoustical Society of America*, 130(3):1475–1487.

Jørgensen, S. and Dau, T. (2013). *Modelling speech intelligibility in adverse conditions.*, volume 787 of *Advances in Experimental Medicine and Biology*, pages 343–351. Springer.

Jørgensen, S. and Dau, T. (2014). *Modeling speech intelligibility based on the signal-to-noise envelope power ratio*. PhD thesis, Technical University of Denmark, Department of Electrical Engineering.

Jørgensen, S., Ewert, S. D., and Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *The Journal of the Acoustical Society of America*, 134(1):436–446.

Kanedera, N., Arai, T., Hermansky, H., and Pavel, M. (1999). On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication*, 28(1):43–55.

Kanedera, N., Hermansky, H., and Arai, T. (1998). On properties of modulation spectrum for robust automatic speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-98)*, volume 2, pages 613–616.

Kay, R. and Matthews, D. (1972). On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones. *The Journal of physiology*, 225(3):657–677.

Kim, C. and Stern, R. M. (2009). Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction. In *Proceedings of Interspeech*, pages 28–31, Brighton, UK.

Kim, C. and Stern, R. M. (2010). Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'10)*, pages 4574–4577.

Kirchhoff, K. and Bilmes, J. A. (2000). Combination and joint training of acoustic classifiers for speech recognition. In *Proceedings of ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*.

Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, 87(2):820–857.

Kolossa, D. and Haeb-Umbach, R., editors (2011). *Robust Speech Recognition of Uncertain or Missing Data: Theory and Applications*. Springer.

Lamel, L. F., Kassel, R. H., and Seneff, S. (1989). Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proceedings of ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases*, pages 100–109.

Lee, D. and Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Lee, K.-F. and Hon, H.-W. (1989). Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648.

Lewicki, M. S. and Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural computation*, 12(2):337–365.

Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):745–777.

Li, X., Singh, R., and Stern, R. M. (2002). Combining search spaces of heterogeneous recognizers for improved speech recogniton. In *Proceedings of Interspeech*.

Lim, J. S. and Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604.

Lippmann, R. (1996). Speech Recognition by Humans and Machines: Miles to go before we sleep. *Speech Communication*, 18(3):247–248.

Luo, H. and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6):1001–1010.

Macho, D., Mauuary, L., Noé, B., Cheng, Y. M., Ealey, D., Jouvet, D., Kelleher, H., Pearce, D., and Saadoun, F. (2002). Evaluation of a noise-robust DSR front-end on Aurora databases. In *Proceedings of Interspeech*, pages 17–20, Denver, Colorado, USA.

Maeda, S. (1982). A digital simulation method of the vocal-tract system. *Speech communication*, 1(3-4):199–229.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2009a). Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, (ICML '09), pages 689–696, Montreal.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. (2008a). Discriminative learned dictionaries for local image analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008).*, pages 1–8. IEEE.

Mairal, J., Elad, M., and Sapiro, G. (2008b). Sparse learned representations for image restoration. In *Proceedings of the 4th World Conference of the International Association for Statistical Computing (IASC)*. Citeseer.

Mairal, J., Ponce, J., Sapiro, G., Zisserman, A., and Bach, F. R. (2009b). Supervised dictionary learning. In *Proceedings of Advances in neural information processing systems*, pages 1033–1040.

Mallat, S. (1999). *A wavelet tour of signal processing*. Academic press.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014a). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010.

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2014b). Mechanisms of noise robust representation of speech in primary auditory cortex. *Proceedings of the National Academy of Sciences*, 111(18):6792–6797.

Meyer, B. T. (2013). What's the difference? Comparing humans and machines on the Aurora-2 speech recognition task. In *Proceedings of Interspeech*, pages 2634–2638.

Meyer, B. T., Brand, T., and Kollmeier, B. (2011). Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes. *The Journal of the Acoustical Society of America*, 129(1):388–403.

Misra, H. (2006). *Multi-stream processing for noise robust speech recognition*. PhD thesis, EPFL, Lausanne, Switzerland.

Mlouka, M. and Liénard, J. (1974). Word recognition based either on stationary items or on transitions. In *Proceedings of 2nd Speech Communication Seminar*, pages 257–263, Stockholm. Almquist & Wiksell International.

Moore, B. (2008). Basic auditory processes involved in the analysis of speech sounds. *Philosophical Transactions of the Royal Society London*, 363:947–963.

Moritz, N., Anemüller, J., and Kollmeier, B. (2011). Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5492 –5495, Prague.

Najnin, S. and Banerjee, B. (2015). Improved speech inversion using general regression neural network. *Journal of the Acoustical Society of America*, 138(3):229 – 235.

Ness, S. R., Walters, T., and Lyon, R. F. (2012). Auditory sparse coding. In Li, T., Ogihara, M., and Tzanetakis, G., editors, *Music Data Mining*, pages 33487–2742. CRC Press, Boca Raton.

Ntalampiras, S., Potamitis, I., and Fakotakis, N. (2011). Probabilistic novelty detection for acoustic surveillance under real-world conditions. *IEEE Transactions on Multimedia*, 13(4):713–719.

Okawa, S., Bocchieri, E., and Potamianos, A. (1998). Multi-band speech recognition in noisy environments. In *Proceedings of International Conference Acoustics, Speech and Signal Processing (ICASSP-98)*, pages 641–644.

Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325.

Olshausen, B. A. and Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487.

Pal, S. K. and Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on Neural Networks*, 3(5):683–697.

Paliwal, K., Schwerin, B., and Wójcicki, K. (2011). Role of modulation magnitude and phase spectrum towards speech intelligibility. *Speech Communication*, 53(3):327–339.

Parihar, N. and Picone, J. (2002). DSR front-end large vocabulary continuous speech recognition evaluation. Technical report, Mississippi State University. Tech. Rep.

Parihar, N., Picone, J., Pearce, D., and Hirsch, H. (2004). Performance analysis of the Aurora large vocabulary baseline system. In *Proceedings of 12th European Signal Processing Conference*, pages 553–556. IEEE.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., and Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS-Biology*, 10(1):e1001251.

Pati, Y. C., Rezaiifar, R., and Krishnaprasad, P. (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, pages 40–44. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos,

A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pham, D.-S. and Venkatesh, S. (2008). Joint learning and dictionary construction for pattern recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*., pages 1–8. IEEE.

Pichevar, R., Najaf-Zadeh, H., Thibault, L., and Lahdili, H. (2011). Auditory-inspired sparse representation of audio signals. *Speech Communication*, 53(5):643–657.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *Proceedings of Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Rabiner, L. R. and Gold, B. (1975). *Theory and Application of Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ.

Raj, B. (2000). *Reconstruction of incomplete spectrograms for robust speech recognition*. PhD thesis, Dept. Electr. & Comput. Eng., Carnegie Mellon Univ., Pittsburgh, PA.

Raj, B., Seltzer, M. L., and Stern, R. M. (2004). Reconstruction of missing features for robust speech recognition. *Speech Communication*, 43:275–296.

Raj, B., Singh, R., and Stern, R. M. (1998). Inference of missing spectrographic features for robust speech recognition. In *Proceedings of ICSLP*, volume 98, pages 1491–1494.

Raj, B. and Stern, R. (2005). Missing feature approaches in speech recognition. *IEEE Signal Processing Magazine*, 22(5):101–116.

Ramirez, I., Sprechmann, P., and Sapiro, G. (2010). Classification and clustering via dictionary learning with structured incoherence and shared features. In

Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), pages 3501–3508. IEEE.

Ramírez, J., Segura, J., Benítez, C., de la Torre, A., and Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 42(3-4):271–287.

Rangachari, S. and Loizou, P. (2006). A noise estimation algorithm for highly non-stationary environments. *Speech Communication*, 28:220–231.

Rennies, J., Brand, T., and Kollmeier, B. (2011). Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet. *The Journal of the Acoustical Society of America*, 130:2999–3012.

Roweis, S. (1998). EM algorithms for PCA and SPCA. *Advances in neural information processing systems*, pages 626–632.

Rubin, P. E., Baer, T., and Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 70:321 – 328.

Saon, G., Padmanabhan, M., Gopinath, R., and Chen, S. (2000). Maximum likelihood discriminant feature spaces. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'00)*, volume 2, pages II1129–II1132. IEEE.

Schädler, M. R., Meyer, B. T., and Kollmeier, B. (2012). Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *The Journal of the Acoustical Society of America*, 131(5):4134–4151.

Seide, F., Li, G., and Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. In *Proceedings of Interspeech*, pages 437–440.

Seung, H. S. and Lee, D. D. (2000). The manifold ways of perception. *Science*, 290(5500):2268–2269.

Shi, G., Shanechi, M. M., and Aarabi, P. (2006). On the importance of phase in human speech recognition. *IEEE Transactions on Audio, Speech and Language processing*, 14:1867–1874.

Singh, R., Seltzer, M. L., Raj, B., and Stern, R. M. (2001). Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'01)*, volume 1, pages 273–276. IEEE.

Smith, L. N. and Elad, M. (2013). Improving dictionary learning: Multiple dictionary updates and coefficient reuse. *IEEE Signal Processing Letters*, 20(1):79–82.

Sroka, J. J. and Braida, L. D. (2005). Human and machine consonant recognition. *Speech Communication*, 45(4):401–423.

Stevens, K. N. (2000). *Acoustic phonetics*, volume 30. MIT press.

Stouten, V., Wambacq, P., et al. (2006). Model-based feature enhancement with uncertainty decoding for noise robust ASR. *Speech communication*, 48(11):1502–1514.

Sun, Y., Cranen, B., Gemmeke, J. F., Boves, L., ten Bosch, L., and Doss, M. M. (2012a). Using sparse classification outputs as feature observations for noise-robust ASR. In *Proceedings of Interspeech*.

Sun, Y., Doss, M. M., Gemmeke, J. F., Cranen, B., ten Bosch, L., and Boves, L. (2012b). Combination of sparse classification and multilayer perceptron for noise-robust ASR. In *Proceedings of Interspeech*, Portland.

Sun, Y., Gemmeke, J. F., Cranen, B., ten Bosch, L., and Boves, L. (2014). Fusion of parametric and non-parametric approaches to noise-robust ASR. *Speech Communication*, 56:49–62.

ten Bosch, L., Boves, L., and Ernestus, M. (2013). Towards an end-to-end computational model of speech comprehension: Simulating a lexical decision task. In *Proceedings of Interspeech 2013: 14th Annual Conference of the International Speech Communication Association*, pages 2822–2826.

ten Bosch, L., Boves, L., Tucker, B., and Ernestus, M. (2015). Diana: towards computational modeling reaction times in lexical decision in north american english. In *Proceedings of Interspeech 2015: 16th Annual Conference of the International Speech Communication Association*, pages 1576–1580.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). Global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Thompson, J. and Atlas, L. (2003). A non-uniform modulation transform for audio coding with increased time resolution. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 397–400, Hong Kong.

Tibrewala, S. and Hermansky, H. (1997). Multi-stream approach in acoustic modeling. In *Proceedings of DARPA Large Vocabulary Continues Speech Recognition, LVCSR-Hub5 Workshop*, pages 1255–1258.

Tipping, M. E. and Bishop, C. M. (1999a). Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482.

Tipping, M. E. and Bishop, C. M. (1999b). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 21(3):611–622.

Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., and Deller Jr., J. R. (2002). Approaches to language identification using gaussian mixture models and shifted delta cepstral features. In *Proceedings of International Conference on Spoken Language Processing*, pages 89–92, Denver.

Tosic, I. and Frossard, P. (2011). Dictionary learning. *IEEE Signal Processing Magazine*, 28(2):27–38.

van Dijkhuizen, J. N., Festen, J. M., and Plomp, R. (1991). The effect of frequency-selective attenuation on the speech-reception threshold of sentences in conditions of low-frequency noise. *The Journal of the Acoustical Society of America*, 90(2):885–894.

Van hamme, H. (2004). Robust speech recognition using cepstral domain missing data techniques and noisy masks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing .(ICASSP'04)*, volume 1, pages I–213. IEEE.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag.

Varga, A. and Steeneken, H. (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251.

Virtanen, T., Singh, R., and Raj, B., editors (2012). *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, Hoboken, NJ.

Wakin, M. B. (2010). Manifold-based signal recovery and parameter estimation from compressive measurements. *arXiv preprint arXiv:1002.1247*.

Walker, S. H. and Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179.

Wei, Z., Wang, X.-J., and Wang, D.-H. (2012). From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *Journal of Neuroscience*, 32(33):11228–11240.

Willmore, B. and Tolhurst, D. J. (2001). Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems*, 12(3):255–270.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227.

Xiao, X., Chng, E. S., and Li, H. (2008). Normalization of the speech modulation spectra for robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1662–1674.

Yang, J., Yu, K., Gong, Y., and Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009).*, pages 1794–1801. IEEE.

Yang, J., Yu, K., and Huang, T. (2010a). Supervised translation-invariant sparse coding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, pages 3517–3524. IEEE.

Yang, M., Zhang, L., Feng, X., and Zhang, D. (2011). Fisher discrimination dictionary learning for sparse representation. In *Proceedings of IEEE International Conference on Computer Vision (ICCV 2011)*, pages 543–550. IEEE.

Yang, M., Zhang, L., Yang, J., and Zhang, D. (2010b). Metaface learning for sparse representation based face recognition. In *Proceedings of 17th IEEE International Conference on Image Processing (ICIP 2010)*, pages 1601–1604. IEEE.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2009). The HTK book (for HTK version 3.4). Technical report, Cambridge University Engineering Department, Cambridge, UK.

Zhang, Q. and Li, B. (2010). Discriminative K-SVD for dictionary learning in face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, pages 2691–2698. IEEE.

Zhang, X., Trmal, J., Povey, D., and Khudanpur, S. (2014). Improving deep neural network acoustic models using generalized maxout networks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*, pages 215–219. IEEE.

Zuev, D. and Moore, A. W. (2005). Traffic classification using a statistical approach. In *Proceedings of International Workshop on Passive and Active Network Measurement*, pages 321–324. Springer.

Zwicker, E., Flottorp, G., and Stevens, S. S. (1957). Critical band width in loudness summation. *The Journal of the Acoustical Society of America*, 29(5):548–557.

Summary

Automatic speech recognition (ASR) refers to the process of converting acoustic features representing speech to a sequence of words. While state-of-the-art ASR systems have achieved high accuracies in quiet environments, their performance drops rapidly when the speech is corrupted by environmental noise, background music or competing speakers. Meanwhile, listening test experiments with noisy speech have shown that, even in the absence of context predictability, human listeners perform much more accurately than ASR systems. This indicates that the human auditory system is able to exploit more cues from the acoustic signals for accurately recognizing noisy speech. In this thesis, we explored different ways to incorporate knowledge about the human auditory system into the design of an ASR system, aiming at improved noise robustness.

In the first study reported in this thesis (Chapter 2), we explored the idea of learning the latent manifold structure of speech signal, in order to deal with with the data loss due to dominance of noise in spectrographic representation of speech. The underlying idea was that the acoustic variations in speech are limited, because they are produced by an articulatory apparatus that cannot change the shape of the vocal tract in arbitrary ways. Therefore, we can assume that speech signals form a (nonlinear) manifold of much lower dimensionality than that of the embedding signal space. We developed a Bayesian model of the speech manifold using a mixture of probabilistic principal component analysers. We obtained a very compact model for the speech manifold, which we employed for reconstructing the noise-dominated elements of speech spectrogram. For this purpose we developed a reconstruction procedure in which maximum likelihood estimates of the missing elements are computed, given the model and the speech-dominated elements in the speech spectrogram. We showed that as long as there is a sufficient number

of reliable measurements in the spectro-temporal representation of a speech signal, the reconstructed signal is accurate enough for the ASR back-end to achieve high recognition performance. We obtained significant improvement in the ASR accuracy on both small and medium size vocabulary speech recognition tasks. We also observed that the recognition performance is relatively insensitive to acoustic variations introduced by external sources, such as noise types and different microphones. This can be interpreted as a confirmation that the manifold model effectively captures the natural variation of speech in the acoustic space.

In the second study (Chapters 3 and 4) we searched for feature representations inspired by models designed to capture the most salient properties of the human peripheral auditory system. The output of the model is the so-called modulation spectrum, which provides a rich and highly redundant feature space to explore. We used exemplar based sparse coding to estimate sub-word unit posterior probabilities from the redundant feature representations, without any need for dimensionality reduction. The sparse coding approach has a conceptual connection with some conjectures about the way concrete tonotopic representations are mapped in the human brain. We conducted a systematic analysis of the impact that different configurations of a modulation frequency filterbank have on the performance of a noise-robust ASR system. We observed that an auditory frontend, which uses multiple bandpass filters to decompose the modulation frequency spectrum, is promising for ASR. To increase the noise robustness of such a system, the features have to be selective in the low modulation frequencies. This feature is similar to human listeners who rely more on time context in very noisy conditions. Moreover, the insight gained in this study suggests that conventional ASR systems might benefit from using multi-range dynamic coefficients.

In our study on human-inspired features, we observed that an exemplar-based ASR system yields lower performance at high SNR levels, compared to recognizers that use supervised learning methods, such as Multi-layer perceptrons, for estimating the posterior probability of sub-word units. This motivated us to include learning in the construction of the dictionary for sparse coding. Therefore, the third study of the thesis (Chapter 5) was on customizing the dictionary learning procedure in such a way that the probabilistic classification accuracy is also considered as an objective during learning. A dictionary learned in that way was expected to provide improved estimated posterior probabilities of speech units in a

sparse classification procedure. For this purpose we proposed class-likelihood consistent dictionary learning, which not only minimizes the reconstruction error of sparse coding, but is additionally constrained in such a way that the sparse codes also provide an accurate estimation of the class membership probabilities for each observation. We showed that the proposed method yields superior performance, both in terms of classification accuracy and in terms of the accuracy with which class likelihoods are estimated. However, the success of the proposed method in speech recognition tasks depends on whether we are able to provide suitable reference class-likelihood vectors and also on whether or not the recognizer back-end is able to benefit from the improved class-likelihood consistency. We continued the study by investigating the extent to which posterior probability estimation based on learned dictionaries increases noise robustness in a sparse coding based ASR system. From the experimental results, we can conclude that in the absence of noise samples to train the classifier with or to form a noise exemplar dictionary from, our proposed learning method will provide a solution which, compared to MLP or an exemplar approach, is more noise robust.

The insights gained from the three studies can be combined to sketch a promising future for noise-robust ASR systems in the absence of the huge amounts of data that are needed for training deep neural networks.

Samenvatting

Automatische spraakherkenning (ASH) is een proces waarin akoestische kenmerken van spraaksignalen gebruikt worden om die signalen om te zetten in een rij woorden. Geavanceerde ASH systemen werken heel goed met spraak die opgenomen is in rustige omgevingen. Maar als de signalen 'vervuild' zijn door omgevingsgeluiden, muziek op de achtergrond of spraak van andere sprekers, dan zakt de nauwkeurigheid waarmee woorden herkend worden aanzienlijk. Experimenten waarin mensen luisteren naar spraak gecombineerd met achtergrondgeluid hebben aangetoond dat mensen dat soort spraak met een veel grotere nauwkeurigheid herkennen dan ASH systemen. Dit suggereert dat het menselijke auditieve systeem meer relevante informatie uit verruiste spraaksignalen kan halen dan de meest geavanceerde ASH systemen. In dit proefschrift hebben we geprobeerd om kennis over het menselijke auditieve systeem te gebruiken bij de ontwikkeling van ASH systemen, met de bedoeling om de herkenprestaties in lawaaierige omgevingen te verbeteren.

In de eerste studie waarvan dit proefschrift verslag doet (Hoofdstuk 2) hebben we onderzocht of de topologische ruimte waarin spraaksignalen gerepresenteerd zijn gebruikt kan worden voor het reconstrueren van die segmenten in de spectrotemporale representatie die overstemd worden door het achtergrondlawaai. Die poging is gebaseerd op het inzicht dat de manier waarop de akoestische kenmerken van spraaksignalen kunnen veranderen in de tijd beperkt wordt door het feit dat die signalen geproduceerd zijn door het menselijke articulatie-apparaat. De fysieke kenmerken van dat apparaat zorgen ervoor dat het niet op een willekeurige manier van vorm, en daarmee van akoestische eigenschappen, kan veranderen. Daarom is het aannemelijk dat de dimensionaliteit van de (waarschijnlijk niet-lineaire) topologische ruimte waarin spraaksignalen bestaan veel kleiner is dan de dimensie van de akoestische observaties. We hebben een Bayesiaans model ontwikkeld dat de

niet-lineaire topologie benadert als een mix van lineaire probabilistische principale componenten analysatoren. Een heel compact model van die topologie maakt het mogelijk om de elementen van spectro-temporele representatie die ontbreken doordat ze overstemd zijn door achtergrondlawaai te reconstrueren. We konden laten zien dat onze methode de spectra van verruiste signalen nauwkeurig kan reconstrueren, mits er een minimaal aantal betrouwbare, niet door achtergrondlawaai overstemde, elementen beschikbaar zijn. De reconstructie van de spectra leidde tot een significante verbetering van de prestaties van een conventioneel ASH systeem voor een klein en voor een middelgroot vocabulaire. Het bleek ook dat de herkenprestaties tamelijk ongevoelig zijn voor verschillen in het achtergrondlawaai en voor het effect van verschillende typen microfoon. Die bevindingen versterken de claim dat onze benadering van de topologische ruimte van de spraaksignalen een goede representatie vormt van de natuurlijke variatie in die signalen.

In de tweede studie (Hoofdstuk 3 en 4) hebben we geprobeerd om een representatie van akoestische spraaksignalen te ontwikkelen die zo goed als mogelijk aansluit bij de werking van het menselijke perifere auditieve systeem. Die representatie is gebaseerd op het modulatiespectrum, dat een rijke en redundante benadering levert van de empirisch bepaalde tonotopische representatie in de hersenstam en in de auditieve cortex. Om de waarschijnlijkheid dat een kort stukje signaal correspondeert met een fonetische eenheid (een klank) te berekenen hebben we gebruik gemaakt van een techniek die die eenheden probeert te benaderen als een som van een klein aantal voorbeeldsignalen (*exemplar-based sparse coding*). Die methode vereist geen voorafgaande reductie van de dimensionaliteit, en garandeert daarmee dat er niet bij voorbaat potentiëel relevante informatie weggegooid wordt. Er zijn ook interessante overeenkomsten tussen de op voorbeelden gebaseerde methode en de manier waarop tonotopische representaties in de hersenen verwerkt worden. We hebben de invloed onderzocht van de manier waarop de modulatiespectra berekend worden, en met name van het aantal filters en de plaatsing van de filters of de frequentie-as in de modulatiefilterbank. Hoewel het gebruik van modulatiespectra altijd een verbetering opleverde ten opzichte van klassieke spectro-temporele representaties is gebleken dat het vooral van belang is dat het oplossend vermogen in de lage frequenties groot is. Dit komt overeen met bevindingen uit luisterexperimenten met mensen, die zich juist en met name in lawaaierige situaties verlaten op continuïteit in de evolutie van spectra over de tijd. De resultaten van deze studie

maken het waarschijnlijk dat toekomstige ASH systemen baat zullen hebben van het combineren van korte- en lange-termijn akoestische kenmerken.

In de tweede studie zagen wij steeds dat *exemplar-based sparse coding* voordeel biedt in lawaaierige situaties, maar dat dat in zekere mate ten koste gaat van de nauwkeurigheid van de herkenning in ruisvrije omgevingen. Dat leidde tot de derde studie (Hoofdstuk 5) waarin we geprobeerd hebben om de voorbeeldsignalen in de zuinige codering te vervangen door bouwblokken die geleerd kunnen worden uit trainingdata, op een zodanige manier dat die bouwblokken voldoen aan twee voorwaarden: ze moeten leiden tot een nauwkeurige benadering van onbekende stukjes spraak, en tegelijkertijd moeten ze leiden tot een nauwkeurige schatting van de waarschijnlijkheid dat een onbekend stukje spraak correspondeert met een van de sub-woord eenheden. Voor dit doel hebben wij een bestaande methode voor het leren van bouwblokken, die alleen gericht was op nauwkeurige benadering, verder ontwikkeld. Op kunstmatige gegenereerde testdata levert onze nieuwe methode zowel een betere classificatie als een betere schatting van de waarschijnlijkheid van de klassen op. In de praktijk hangt de kracht van onze methode sterk af van de betrouwbaarheid waarmee de gelijkenis en de verwarbaarheid van de klassen (in ons geval de sub-woord eenheden) in de trainingdata geannoteerd kunnen worden. Voor toepassingen in ASH is het bovendien van belang tot op welke hoogte de decoder kan profiteren van nauwkeurigere schatting van de waarschijnlijkheid van de sub-woord eenheden. Uit onze experimenten blijkt dat de nieuwe manier om bouwblokken te leren vooral voordelen biedt in situaties waarin het niet mogelijk is om voorbeelden van spraaksignalen aan te vullen met voorbeelden van relevante ruissignalen.

Door de inzichten uit de drie studies te combineren wordt het mogelijk om nieuwe ASH systemen te ontwikkelen die robuust zijn tegen achtergrondlawaai, zonder daarbij een beroep te hoeven doen op de gigantische hoeveelheden trainingdata die nodig zijn voor de diepe neurale netwerken.

Curriculum Vitae

Sara Ahmadi (November 1983), received the BSc degree in electrical and electronic engineering from Amirkabir University of Technology, Tafresh campus, Iran in 2006. In 2009, she received the MSc degree in electronics engineering from Amirkabir University of Technology, Tehran, Iran. In September 2009 she started as a Ph.D. student on speech processing in the Speech Processing Research Laboratory, Electrical Engineering Department, Amirkabir University of Technology. Later in 2014 she joined a Marie Curie training network at the Centre for Language Studies, Radboud University, Nijmegen, the Netherlands. She then became a joint PhD student at Radboud University and Amirkabir University of Technology. Her main areas of research interest are digital signal processing, automatic speech recognition, statistical modeling, machine learning, pattern classification and sparse coding.

SIKS Dissertation Series

2011
01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models
02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language
03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems
04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference
05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.
06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage
07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction
08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues
09 Tim de Jong (OU), Contextualised Mobile Media for Learning
10 Bart Bogaert (UvT), Cloud Content Contention
11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective
12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining
13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling
14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets
15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval
16 Maarten Schadd (UM), Selective Search in Games of Different Complexity
17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness
18 Mark Ponsen (UM), Strategic Decision-Making in complex games
19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles
20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach
21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems
22 Junte Zhang (UVA), System Evaluation of Archival Description and Access
23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media
24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics
26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns
28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure
29 Faisal Kamiran (TUE), Discrimination-aware Classification
30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions
31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science
33 Tom van der Weide (UU), Arguing to Motivate Decisions
34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training
36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach
37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference
38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization
39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games
40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development
41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control
42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution
43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge
44 Boris Reuderink (UT), Robust Brain-Computer Interfaces

45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection
46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression
48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

---

2012 01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda
02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories
04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications
05 Marijn Plomp (UU), Maturing Interorganisational Information Systems
06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks
07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories
09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms
10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment
11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems
13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
18 Eltjo Poort (VU), Improving Solution Architecting Practices
19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution
20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
26 Emile de Maat (UVA), Making Sense of Legal Text
27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
28 Nancy Pascall (UvT), Engendering Technology Empowering Women
29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval
30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications

35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41 Sebastian Kelle (OU), Game Design Patterns for Learning
42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43 Withdrawn
44 Anna Tordai (VU), On Combining Alignment Techniques
45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering
51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

2013 01 Viorel Milea (EUR), News Analytics for Financial Decision Support
02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics
04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
07 Giel van Lankveld (UvT), Quantifying Individual Player Differences
08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services
12 Marian Razavian (VU), Knowledge-driven Migration to Services
13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling
20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
22 Tom Claassen (RUN), Causal Discovery and Logic
23 Patricio de Alencar Silva (UvT), Value Activity Monitoring
24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning
27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance
28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience

29 Iwan de Kok (UT), Listening Heads
30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support
31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications
32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development
33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere
34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search
35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction
36 Than Lam Hoang (TUe), Pattern Mining in Data Streams
37 Dirk Börner (OUN), Ambient Learning Displays
38 Eelco den Heijer (VU), Autonomous Evolutionary Art
39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems
40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games
41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning
43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

2014 01 Nicola Barile (UU), Studies in Learning Monotone Models from Data
02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method
03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions
04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation
05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability
06 Damian Tamburri (VU), Supporting Networked Software Development
07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior
08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints
09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
10 Ivan Salvador Razo Zapata (VU), Service Value Networks
11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support
12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control
13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains
14 Yangyang Shi (TUD), Language Models With Meta-information
15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria
17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations
19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
24 Davide Ceolin (VU), Trusting Semi-structured Web Data
25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
26 Tim Baarslag (TUD), What to Bid and When to Stop
27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
28 Anna Chmielowiec (VU), Decentralized k-Clique Matching

29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
30 Peter de Cock (UvT), Anticipating Criminal Behaviour
31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data
33 Tesfa Tegegne (RUN), Service Discovery in eHealth
34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.
35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach
36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying
38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.
39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital
40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education
41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text
42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models
43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments
44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.
45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach
46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity
47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

2015 01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response
02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls
03 Twan van Laarhoven (RUN), Machine learning for network data
04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments
05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding
06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis
08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning
11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins
12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
13 Giuseppe Procaccianti (VU), Energy-Efficient Software
14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation
16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners
20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination
21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
22 Zhemin Zhu (UT), Co-occurrence Rate Networks
23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage
24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection
26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure
27 Sándor Héman (CWI), Updating compressed colomn stores

28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO
29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning
31 Yakup Koç (TUD), On the robustness of Power Grids
32 Jerome Gard (UL), Corporate Venture Management in SMEs
33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources
34 Victor de Graaf (UT), Gesocial Recommender Systems
35 Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VU), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UVA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UVA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (UvT), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

---

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UVA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TUE), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UVA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VU), Logics for causal inference under uncertainty
23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VU), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30 Wilma Latuny (UvT), The Power of Facial Expressions
31 Ben Ruijl (Uvt), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VU), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications: User-centered Models for Sharing Location in the Family Life Domain
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR